سازمان مدیریت و برنامه‌ریزی کشور
دفتر انتشارات علمی و مدارک تخصصی

وزارت علوم، تحقیقات و فناوری
مرکز اطلاعات ومدارک علمی ایران

پروژهٔ ایجاد پایگاه مجازی مطالعات طرحها و نقشه های عمرانی کشور
(آرشیو چیذر)
قرارداد شمارهٔ ۱۷۰۶/۲۱۸۹۰۲ تاریخ ۸۱/۱۱/۱۹

طراحی سازمان، فرایندها و استانداردهای خدمات اطلاع‌رسانی پروژه‌های عمرانی کشور

طراحی فرایند منطقی تولید اصطلاحنامه

سیروس علیدوستی
فخرالسادات محمدی

گزارش مرحلهٔ دوم
۱۳۸۳

شماره ثبت: ۵۱۴
تاریخ: ۱۳۹۲/۹/۰۹

طراحی سازمان، فرایندها و
01XF0000000000514
پژوهشگاه علوم و فناوری اطلاعات ایران

طراحی فرایند منطقی اصطلاحنامه

فعالیت های انجام شده برای طراحی فرایند منطقی تولید اصطلاحنامه به قرار زیر است:

مرحله اول شامل تعریف موضوع، دامنه، و اهداف این طرح

مرحله دوم جستجو درپایگاههای اطلاعاتی برای تهیه منابع پیرامون موضوع بوده وسعی شده جستجو با سرفصل‌های مطرح شده در طرح هماهنگ باشد.مرحله بعدی پس از جستجومطالعه منابع انتهای مقالات است که از طریق آن امکان دسترسی به منابع بیشتر(کتاب ، مقاله، گزارش، استاندارد و...) ممکن شد.نتایج جستجو در انتهای این گزارش ضمیمه خواهند شد.

قسمت عمده جستجوها در پایگاههای اطلاعاتی زیر صورت گرفته است:

www.sciencedirect.com

www.emeraldinsight.com

http://vnweb.hwwilsonweb.com

www.proquest.com

www.ebsco.com

www.ingenta.com

www.dois.mimas.ac.uk

فرایند تدوین اصطلاحنامه

سازماندهی اطلاعات

نمایه سازی

نمایه سازی همارا

نمایه سازی پیش همارا

نمایه سازی پس همارا

زبان نمایه سازی

زبان طبیعی

زبان کنترل شده / مهارشده

اصطلاحنامه ها

اهداف

تعاریف

تفاوت اصطلاحنامه با فرهنگ لغات

تفاوت اصطلاحنامه با سرعنوان موضوعی

انواع

ازلحاظ کاربرد

کلاسیک

نمایه سازی

جستجو

از لحاظ شمول زبانی

تک‌زبانه

چندزبانه

نمایش اصطلاحات

چاپی

الفبایی

سلسله‌مراتبی

رده ای

گردان

گرافیکی

درختی

پیکانی

نمایش پیوسته(online)

الفبایی

سلسله‌مراتبی

رده ای

گردان

گرافیکی

درختی

پیکانی

استاندارد های تدوین اصطلاحنامه

اصطلاحنامه تک زبانه

استاندارد بین المللی (ISO2788)

استاندارد های ملی

استاندارد بریتانیا(BS6723)

استاندارد آمریکا(ANI/NISO Z39.19)

استاندارد فرانسه(ANFOR NFZ47-100)

استاندارد آلمان(DIN1463)

اصطلاحنامه چندزبانه

استاندارد بین المللی (ISO5964)

استاندارد ملی

استاندارد بریتانیا(BS6723)

تدوین اصطلاحنامه

اصطلاحنامه تک زبانه

تعریف حوزه موضوعی

محدوده اصلی(هسته)

محدوده فرعی(حاشیه ای)

ویژگی و شکل اصطلاحنامه

ساختار

سطح پیش همارایی

اخص بودن اصطلاح

طراحی اصطلاحنامه

روش قیاسی

روش استقرایی

انتخاب اصطلاحات

اصطلاح مرجح

بسامدی قابل قبول

ضمانت ادبی

ضمانت استفاده

نکات دستوری

شکل مفرد یا جمع

اصطلاح علمی / عامیانه

یادداشت دامنه

همنویسه ها

توضیحگر داخل پرانتز

تعیین روابط

رابطه معادل / هم ارز

مترادف‌ها

رابطه‌سلسه مراتبی

کل وجز

جنس و گونه

رابطه موردی

تعیین رده ها

اصطلاحنامه دو زبانه

اصطلاحنامه فارسی_انگلیسی

ترجمه اصطلاحنامه انگلیسی به فارسی

تدوین اصطلاحنامه فارسی_انگلیسی

منابع تدوین اصطلاحنامه

منابع چاپی

واژه نامه های تخصصی فارسی

واژه نامه های آخرمدارک تخصصی

طرح های رده‌بندی

سرعنوان موضوعی

اصطلاحنامه‌های موجود وفهرست های اصطلاحات

منابع غیرچاپی

بانک‌های اصطلاح شناسی

فهرست واژگان بانک های اطلاعاتی

اصطلاحات ثبت شده در پرسش نامه و پرونده کاربران

دانش و تجربه تدوین کننده

تعیین توصیفگرفارسی درمقابل انگلیسی

تعیین مترادف ها وشبه مترادف‌ها

هم نویسه‌ها درزبان فارسی

درج توضیحگر داخل پرانتز

رعایت ضوابط دستوری

رعایت یک دستی در رسم الخط و املاحروف

حروف‌نگاری یا آوانگاری کوته‌نوشته‌ها و سرواژه ها

تکوین(شکل گیری تدریجی)اصطلاحنامه فارسی_انگلیسی

نمایه‌سازی مدارک فارسی با استفاده از اصطلاحنامه‌های مستند انگلیسی

تدوین ضوابط و معیارها(دستوری، املایی، نوشتاری)

تهیه فهرستی از اصطلاحنامه‌های معتبر تخصصی

اولویت‌بندی اصطلاحنامه‌ها در هر زمینه موضوعی

استخراج کلیدواژه از مدارک

تعیین توصیفگر انگلیسی(با ذکر ماخذ به صورت نشانه)

تطبیق بین اصطلاحنامه‌ها

تعیین توصیفگر فارسی

یکپارچگی بانک های واژگان یک سازمان/ یک حوزه موضوعی

تعیین مترداف ها و شبه‌مترادف‌ها

هم‌نویسه‌ها در زبان فارسی

توصیفگر داخل پرانتز

شکل‌گیری هم زمان اصطلاحنامه

انتخاب یک اصطلاحنامه کلان

اقتباس ساختار و تقسیمات کلی موضوعی

دسته بندی کلی توصیفگر های فارسی در رده‌های کلی موضوعی

تعیین زیر رده‌ها

برقراری رابطه ها

تهیه یادداشت‌دامنه برایی توصیفگر

کنترل ارجاعات

مدیریت اصطلاحنامه

کنترل

ثبت تاریخ افزوده(added date)

یادداشت موقعیت اصطلاح (posting note)

روزآمد کردن

بررسی اعتبار اصطلاح (پویایی زبان)

تاریخ حذف

بررسی اعتبار اربطه ها (گسترش علوم)

کاربر پسند بودن اصطلاحنامه

بازیابی اطلاعات

ابزارهای بازیابی

اصطلاحنامه جستجو

منابع و ماخذ

PERGAMON

Information Processing and Management 38 (2002) 363–379

INFORMATION PROCESSING & MANAGEMENT

www.elsevier.com/locate/infoproman

Exemplary documents: a foundation for information retrieval design

David C. Blair [a,*], Steven O. Kimbrough [b]

[a] *Computer and Information Systems, Graduate School of Business, University of Michigan, Ann Arbor, MI 48109-1234, USA*

[b] *The Wharton School, University of Pennsylvania, 1322 Steinberg Hall-Dietrich Hall, Philadelphia, PA 19104-6366, USA*

Received 8 April 2000; accepted 19 January 2001

Abstract

Documents are generally represented for retrieval by either extracting index terms from them or by creating and selecting from an external set of candidate terms. There are many procedures for doing this, but while work continues along these dimensions, there have been relatively few attempts to change this basic process. Of particular importance is the creation of indexing schemes for retrieval systems in non-library contexts. Here, the cost of developing an indexing scheme independent of the documents to be retrieved is often considered too high to implement. As a result, simple full-text retrieval or, to a lesser extent, automatic extractive or associative indexing methods are the predominant methods used in non-library contexts. This paper suggests an alternative document representation method based on what we call *exemplary documents*. Exemplary documents are those documents that describe or exhibit the intellectual structure of a particular field of interest. In so doing, they provide both an indexing vocabulary for that area and, more importantly, a narrative context in which the indexing terms have a clearer meaning. Further, it is much easier to develop an indexing scheme by using exemplary documents than it is to do so from scratch. © 2002 Elsevier Science Ltd. All rights reserved.

1. Introduction

Information, or document, retrieval is no longer just the library problem, and both the designers of management information systems and commercial retrieval system providers are coming to see this. Unlike libraries, organizational document retrieval systems are often rapidly

* Corresponding author.
E-mail address: dcblair@umich.edu (D.C. Blair).

0306-4573/02/$ - see front matter © 2002 Elsevier Science Ltd. All rights reserved.
PII: S0306-4573(01)00027-9

built, do not follow traditional library methods of organization and representation, and frequently involve "mission-critical" information – information for which there may be a substantial penalty if it cannot be retrieved in a timely fashion. Certainly there are many non-trivial challenges to be met if we are to build effective information retrieval systems, but one design factor is critical: the level of effectiveness that an information retrieval system can attain is crucially dependent on how the documents in a collection are represented (Blair, 1990). No computerized retrieval system or retrieval algorithm, however complex, can adequately compensate for documents that are represented poorly or inappropriately. But the widespread need for non-library document retrieval systems which require high levels of retrieval effectiveness has meant that system designers are finding themselves dealing with document collections that may have no obvious representation scheme that could provide access to the intellectual content of those documents. It had always been hoped that some underlying universal representation scheme could be discovered that might be used to classify all information or knowledge. But attempts to find such universal classification schemes are as old as Platonic Realism, and the single, timeless classification of knowledge remains elusive. The thoughtful system designer, faced with the demands of building mission-critical information retrieval systems, is soon confronted with the fact that there is not one, ideal classification scheme for textual information, but myriad ones – uncountable, dynamic and ad hoc. This phenomenon was demonstrated most convincingly in Swanson's important, but not widely circulated study (Swanson, 1966; discussed in Blair, 1990). Swanson found that when indexers were encouraged to assign as many index terms as possible to a document, they could come up with a virtually unlimited number of reasonable descriptions of the intellectual content of those documents – and in some cases, the total number of descriptions for a document was even greater than the length of the document being indexed. The reasons for this phenomenon are not entirely clear, but certainly the inherent creativity or productivity of natural language is, in part, responsible (Blair, 1990, pp. 170–171, Langendoen & Postal, 1984). It is also the case, that even when a reasonably good representation scheme is put together the process is so labor-intensive that only the most persistent systems designer can carry it out; and even a well-constructed representation scheme suffers from an important trade-off: the more complex and finely tuned the system of representation is, the less comprehensible that system may be for the average searcher. Clearly, the challenges of designing a representation scheme for a set of documents where no obvious representation scheme already exists and high levels of retrieval effectiveness are needed, are formidable.

2. The simple full-text retrieval solution

Faced with the significant challenges of formulating an intellectual structure for a collection of documents where no such structure has existed before, the designer of document retrieval systems often utilizes a simple full-text retrieval system. In such a system, the entire text of the collected documents is stored on line and the searchers retrieve documents simply by guessing the words and phrases that might occur in the documents that they want (but do not occur in the documents they do not want). While there are complex versions of this search model, what we are discussing here is the most simple version of this model where the search process is left entirely up to the searcher and is not supplemented by any automatic techniques. This is by far the most common

version of commercial full-text retrieval systems, and is a search option for a majority of computerized document retrieval systems (Delphi Consulting Group Inc., 1993). Of course, this kind of system relieves the system designer from the task of constructing an intellectual framework for the documents by accepting the actual text of the documents as a default representation scheme (non-content words, such as articles and prepositions, as well as duplicates of the words in text are generally not used to represent the documents). Such simple full-text systems, while easy to set up from an intellectual point of view, do not furnish very high levels of recall when providing access to a document collection of realistic size. In short, they trade off lower indexing and design effort for higher search effort. Blair and Maron (1985, 1990) demonstrated a *maximum* recall value of 20% for such a system in an operational setting. In spite of this, there are members of the information retrieval research community who still advocate the use of document retrieval systems with such low performance levels. In response to Blair and Maron (1985), Salton remarked, "...not only is this level of performance (i.e., 20% recall) typical of what is achievable in existing, operational retrieval environments, but that it actually represents a *high order* of retrieval effectiveness." (Salton, 1986, p. 649. Author's emphasis.) Yet, increasingly, recall levels of 20%, or less, are not seen as acceptable for operational systems, especially in critical applications outside the walls of university libraries and experimental test collections. In more and more current information retrieval applications, the consequences of low levels of access can be disastrous:

> The effects of a failure of 'document control' can be dramatic. A major utility company was required to shut down four nuclear reactors because of lost repair instructions (at a loss of $2m per day). The US department of defense estimates that half of all military accidents result from missing or inaccurate technical information. A major airline was fined $10 k per take-off because of out-of-date maintenance information. A major drug company lost its entire R&D investment owing to inability to provide timely documentation (Fleischer, 1991)

If information retrieval researchers truly believe that a 20% recall level represents a "...*high order* of retrieval effectiveness..." then they are certainly out of touch with the demands of the market-place.

Efforts to design more effective operational information retrieval systems are certainly proceeding. Commercial systems such as Queria or Thunderstone's "Metamorph", among others, are developing and implementing automatic ways of improving searches based on the intellectual content of documents. But another way to improve access to documents is to investigate better ways of representing the intellectual structure and content of those documents. It is in this area that this paper aspires to contribute. Efforts in this direction are no less promising than automatic techniques, and improvements in both areas would be complementary. But progress in this latter area is uncertain – as we mentioned previously, efforts to identify an underlying knowledge structure to map all possible document representations onto have proved difficult and problematic. What is needed are some general principles that would help us to focus on which ways of representing documents are better than others. One suggestion is to relate the representation scheme for a document collection to the activities that use or generate the documents in the collection (Blair, 1990). This paper is an attempt to look at another (though related) basis for the representation scheme of a document collection – *exemplary documents*. (N.B. This section should not be construed as saying that the use of exemplary documents in IR systems will produce recall

values greater than 20%. Without empirical studies, it would be impossible to make such estimates, given our present meager ability to model IR system performance. It may be, too, that the major benefit of exemplary documents will not be an improvement in recall, but a speed-up or facilitation of the indexing and design process.)

3. Exemplary documents

In many document collections some documents are more representative of the intellectual structure of that collection than others are. That is, the intellectual structure of the document is similar (in whole or in part) to the intellectual structure of the discipline or field of study of which the document is a part. These are what we call "exemplary documents." Exemplary documents provide not only a view of the issues discussed in many of the other documents in the collection, but they also provide a framework which often links those issues together into a more meaningful whole. Further, they also provide the vocabulary in which those issues are discussed. In essence, exemplary documents provide an "intellectual road map" to the semantic content of a given set of documents. Many document collections, then, can be divided into two kinds of documents: the documents to which access is desired, and a much smaller set of exemplary documents which may be used to build an intellectual structure on which some or all of that access can be based. If these exemplary documents do exist in sufficient numbers, and it is our contention that they do, it may be possible to exploit them as an organizing principle for a collection of documents.

4. Exemplary documents: examples

So far, we have said that exemplary documents provide a conceptual "view" of the issues and topics that a set of documents is concerned with. While helpful, such a rough definition must be made more precise if we are to use it to identify actual exemplary documents. Perhaps the best way to do that would be to begin by discussing several examples of exemplary documents.

4.1. Research literature

The most obvious kind of exemplary document is the "survey article" that appears in the literature of a particular research area. This kind of article explicitly discusses the major issues or topics in a field of study, and often provides specific references from these issues to significant publications, authors and institutions that are important in this area. But survey articles do not just provide a "topography" of the intellectual landscape of the field, they also discuss the issues of the field *using the professional vernacular of that field*; that is, they are demonstrations of *how* the field refers to, and talks about, these issues and topics. This is not to say that the language of that field is fixed. There are few active fields in which that would be the case. But insofar as there might be preferred or popular ways of discussing the topics and issues of a field, those linguistic preferences would likely be exemplified in the survey article. In this way, the survey article can act as a kind of conservative linguistic process that demonstrates how, out of the many ways a topic *could be* discussed, it usually *is* discussed in that field of inquiry. Survey articles, therefore, provide

several essential elements for representing the intellectual content of documents in a particular field:

- A description of the major issues or topics in the field, or part of the field (often including those issues that are in direct conflict with each other).
- A demonstration of the language used to refer to the issues and topics with which the field is concerned.
- References to some of the significant authors/researchers in the field.
- A description of major techniques, methodologies, ways of working, etc. (where applicable).

N.B. Obviously the citations given in the survey article could also be followed up to find the names of institutions and journals that also figure prominently in that field.

Survey articles are not the only exemplary documents in research literature, but they are usually the easiest kind of exemplary document to identify. Those exemplary documents which are not survey articles share with the survey article the ability to describe the major issues and topics of a field. Such documents might be editorials or opinion papers written by prominent members of a field's research community in which they might draw on their experience in the field to reflect on its conceptual or intellectual structure. This kind of exemplary document tends to be more abstract than a survey article. Ostensibly, it contains the views of a single individual, while the survey article is frequently a synthesis of the work of many individuals. But since the author of an editorial or opinion paper is usually a prominent member of the research community, his or her views, while personal, still often represent a consensus of sorts.

There is also a hybrid exemplary document that can have characteristics of both the survey article and the reflective opinion paper. This is the lead article (usually written by a guest editor) in a special issue of a research journal devoted to a specific topic. These lead articles vary widely in content, and in many cases they may not qualify as exemplary documents at all. But in some instances, such articles may both survey a particular field, or sub-field, as well as propose an abstract intellectual framework in which to fit the topics and issues of that subfield.

The idea of an exemplary document recalls Kuhn's notion of an "exemplar" (Kuhn, 1970, p. 186ff), so it may be useful to take a moment to describe the relation, if any, between exemplars and exemplary documents. For Kuhn, the principal activity of normal science is puzzle-solving. Classic puzzles that a field has already solved serve as what Kuhn calls exemplars. These exemplars are the most important component of the disciplinary matrix that constitutes the paradigm of a scientific field. They serve as good examples of the kind of puzzles members of that field should work on, and thereby also become models for the kind of new puzzles members of that field should devote themselves to. Further, they also have a pedagogical purpose: the student becomes a full-fledged member of that field by working through and solving these exemplary puzzles. If he does enough of them, he will attain not only the accepted puzzle-solving skills and methods of that field, but will also share a similar point of view towards the subject matter and activities of the field (what Kuhn called a "group-licensed way of seeing" (op.cit. p. 189)). This common point of view will allow the neophyte to select the right new puzzles to work on in his/her field. Anyone who has taken an introductory physics laboratory has had first hand experience with this process. In fact, Kuhn insists, it is easy to identify the exemplars of any stable field – just examine a set of textbooks of the field and see which puzzles they have in common. Those that they have in common are the exemplars of the field.

Although exemplary documents are not puzzles, they *do* share some characteristics of Kuhn's exemplars. Specifically, exemplary documents can also have a pedagogical role in most disciplines. Any document that presents the intellectual structure of a field in a clear and unambiguous way, as many survey articles do, would be an important source from which students could learn the intellectual structure of that field. It would help them to attain the group-licensed way of seeing that Kuhn's exemplars do.

Just as the textbooks of a field are important sources for Kuhn's exemplars, they might also have a role to play similar to exemplary documents. Specifically, the organization of a textbook is likely to be similar to the intellectual organization of the field it represents. The organization of the textbook, then, exemplifies the intellectual organization of the field. Further, the vocabulary used in the textbook should be a good model for the indexing vocabulary needed to represent documents in that field. In short, textbooks may provide the same kind of intellectual organization of a field that exemplary documents do. There are a few differences, though. Textbooks are invariably longer than exemplary documents, so they may serve as less concise access mechanisms than exemplary documents do (although, individual chapters of the textbook might be taken separately as exemplary documents that describe the intellectual structure of a part of the discipline). Also, they may represent a more conservative or established view of the field – they describe the field as it is taught, and would probably not represent newer or more controversial areas of the field. Survey articles, on the other hand, would be more likely to describe newer, less established or less widely accepted work in the field. Further, textbooks generally cover more introductory levels of the field than survey articles do since, of course, they are intended to be used by students. Survey articles, on the other hand, often deal with more advanced concepts and are often intended primarily for established members of the field (or, "aspiring" members of the field, such as graduate students). Finally, because of the textbook's need to represent an accepted view of the field it deals with, it may represent an older, less current view of that field. Survey articles can be published more quickly than textbooks, so they can establish an immediacy unattainable in the longer, more consensus-oriented format of the textbook.These caveats aside, textbooks are certainly candidates for exemplary status.

There is another kind of article in the research literature that may bear some relationship to exemplary documents: the seminal paper. While it is possible that an exemplary document may also be a seminal paper in a given field, it is not the case that every seminal article is an exemplary document. An example should make this clearer. As a recent survey makes clear (Krishnan, 1993), the article by Will (1975) is considered seminal in the field of Model Management since it was instrumental in motivating researchers to begin work in this area. But the article itself, while *seminal*, is not *exemplary* since it offers no enduring guide to the major issues, topics or methods of the field as it developed. In the field of information retrieval, Vannevar Bush's paper "As we may think" (Bush, 1990), is frequently mentioned as a seminal paper in the field and has even found a recent relevancy to hypertext-based information retrieval systems. While anticipating hypertext-type document arrangements, Bush's article, of course, does not anticipate how the intellectual structure of hypertext-based information retrieval has recently evolved. Like the Will article (*supra*), it offers no enduring guide to the issues, topics or methods of the information retrieval field, or even the narrower field of hypertext-based retrieval systems, and is thus *not* an exemplary document.

4.2. Large-scale litigation

Corporate and government litigation are just two areas where there is a dramatically increasing number of large-scale lawsuits – lawsuits in which attorneys may need to have access to thousands, or sometimes millions of documents. Further, the successful conduct of the lawsuit is critically dependent on effective document access. The intellectual structure of the information germane to a particular lawsuit is best exemplified in a document called the "complaint". The complaint identifies, very precisely, the issues or events being litigated, the individuals involved in the case, and, if relevant, the institutions involved also. In addition, the complaint will often identify the *chronological sequence of events* that tie all the individuals, issues and institutions together, and will often assert *causal links* between certain key events.

Another kind of exemplary document in litigation consists of the "depositions" given during the pre-trial phase of the lawsuit. These depositions are often explicit narratives given by the major participants in the lawsuits in which they answer questions and discuss the actionable events in question. While the complaint gives the legal (i.e., lawyers') view of the suit, the depositions give the individual participants' views of the events in question. Depositions and complaints often have a narrative structure – they tell a story. The narrative is not only a good way to organize the issues of a lawsuit, but may, in fact, be a fundamental way in which we structure information. To some researchers in cognitive science or related fields (e.g., Bruner, 1990; Turner, 1996), the narrative is one of the central organizing principles of our experience and knowledge. Hence, the deposition may directly model the way we understand the issues of a lawsuit making its use as an organizing principle in an information system even more compelling.

4.3. Public policy

The formulation and analysis of public policy comprises an interesting body of textual information. The literature itself often has a structure that organizes around explicit arguments or debates. Frequently, in this kind of literature, documents will occur which describe the issues and structure of particular debates. For example, we have been building a prototype information retrieval system based on exemplary documents in the area of "technology policy". The debate here revolves around the United States' declining world competitive strength in science and technology. As we gathered documents on technology policy it soon became clear that one particular document exemplified the major issues in this debate, and asserted several causal connections that were the basis for much of the debate in this area. This document was "The competitive strength of US industrial science and technology: strategic issues" (National Science Board., 1992). It satisfied our loose criteria for an exemplary document because it gave a clear presentation of the major issues in the debate over technology policy. It also described the major causal assertions of the debate and cited supporting literature for its view of the debate.

As we looked further into the literature of this debate, we found a second exemplary document – "Technology: The engine of economic growth. A national technology policy for America", authored by the 1992 Clinton/Gore Presidential campaign committee (Clinton/Gore, 1992). Like the National science Board Report, the Clinton–Gore document provided a broad view of the debate and made several causal assertions of its own. But there were also differences between the NSB report and the Clinton–Gore document. Both represented somewhat different views of

the same debate, so there was not a one-to-one correspondence between the issues and assertions of the two documents. Is this a problem? In other words, is there a difficulty if two exemplary documents make different – even opposing – assertions? Further, is there a problem if the assertions of an exemplary document prove incorrect? The answer to each of these questions is "not necessarily". What is essential about exemplary documents is not whether their assertions are correct or incorrect, but whether they can provide an intellectual road map to a body of literature. So, for example, an article that passionately advocates a "creationist" view of evolution might be useful for organizing not only the creationist literature, but also the more "scientific" literature of evolutionary theory. This is because while the creationist exemplary document may argue for the creationist point of view, it may also mention and attack the scientific paradigm for evolution theory, thereby providing a schema for both the scientific and creationist views of the field.

4.4. Project management

Long term projects are examples of information intensive activities that can generate a large number of documents, drawings, engineering specifications, etc., all of which are relevant to the conduct of the project and many of which are essential. Important organizing principles of such document collections would be the chronology of the project and the engineering breakdown of the components or systems which comprise the project. Exemplary documents for project document control would most likely consist of project initiation or enabling documents (such as the original contract specifications), major planning and review documents, documents prepared for major decision points in the course of the project, or evaluations of project checkpoints.

We applied these principles in the development of a document-oriented project management system for the Coast Guard as part of the KSS project (see Kimbrough, Clark, Michael, & Hemant, 1990, for early work in this area). Here we identified an important document, OMB Circular A-109, which defines the major systems acquisitions process for the US Government. This document describes the acquisition process in detail and specifies key subsequent documents that must be created and approved as part of the process. In essence, it prescribes how the activity must progress and which specific subsequent documents must be created to enable the process to continue. Some of these subsequent documents may be exemplary documents themselves. In this case, the primary exemplary document is not hard to identify. In fact, it is a sine qua non of the project management process itself.

The above examples illustrate the kinds of exemplary documents one might find in several different fields. It would not be hard to give similar examples for other information-intensive activities such as corporate strategy, archival management or the management of museum collections.

5. Exemplary documents: characteristics

Looking at the above examples of exemplary documents, we can start to see some of the characteristics that many of them have:

(i) They always provide a synoptic view or survey of at least some of the major topics and issues of a field, or part of a field.

(ii) They often discuss these topics and issues with a vocabulary that is a restricted subset of all the possible ways these issues could be discussed. An exemplary document is a demonstration of the "dialect" of natural language that is used to talk about issues in that field.
(iii) They often indicate the structure or framework of the field by showing, explicitly or implicitly, the relationship between some or all of the issues or topics they identify (e.g., that certain issues are related *causally*; or, how specific events, such as a series of experiments, are related chronologically).
(iv) They often provide explicit bibliographic links to individuals, authors, institutions, documents, etc., that figure prominently in a given literature.
(v) The truthfulness of the assertions made by an exemplary document do not necessarily affect its usefulness as a tool for organizing a body of literature.
(vi) In some, perhaps rare, instances an exemplary document can even *drive* a field's literature; that is, an exemplary document may be so persuasive about identifying the issues and structure of a field that subsequent papers that are written in the area may respond specifically to the issues and their relation spelled out in that exemplary document. A clear historical example of this is the mathematician Hilbert's early 20th century paper that outlined what he believed were 23 major puzzles of mathematics for that time (Hilbert, 1901). His paper became a blueprint for a great deal of the subsequent work in theoretical mathematics for decades afterwards. In our own discussion, exemplary documents for the literature of public policy debate often drive the subsequent literature. In cases like this there may be some correlation between citation rates and exemplary documents, though high citation rates alone would not, we think, identify exemplary documents.

6. Exemplary documents: internal structure

The role of the internal structure of an exemplary document is important enough to discuss in some detail. Such a structure may be explicitly presented in the document, or it may be implicit. There is also a number of ways that the issues identified by an exemplary document can be related: in the complaint of a lawsuit there may often be an explicit assertion of causal relations between specific actionable issues; there may also be a chronological sequence of events that is significant, too. In a paper which surveys the research of a field, the chronological sequence of that research may be a very important part of the structure of that field since it may be used to establish precedent and dependence. The research methods of a given field may also provide an implicit structure for the issues in that field, and these structures may vary from field to field insofar as their respective research methods vary.

There is also another kind of internal structure inherent in exemplary documents that should not be overlooked. This structure is the narrative structure of the document text itself. It almost always contains sections and headings (some survey articles, such as those in *computing surveys*, even have tables of contents for *each* article in an issue). These section headings and divisions offer not only an explicit structuring of the article, they also may offer, implicitly, a structuring of the issues discussed in the article, and this implicit structuring of the article may be used to organize other documents in the field.

On a more subtle level, there is also a structure to the discourse of the document. This discourse may be classifiable, for example, according to categories of Speech Acts (Austin, 1962; Searle,

1969) or "implicatures" (Grice, 1989), and these categories, in turn, imply a structure that may be useful for arranging the issues of a field (see Blair, 1992). The representation of the intellectual content of documents is, fundamentally an issue of language and meaning.

7. Exemplary documents and activity-based document representation

In the first part of this paper, we mentioned that the notion of exemplary documents offered an alternative to representing documents based on the activities that "use or generate the documents in the collection" (Blair, 1990). On closer examination, however, it becomes clear that exemplary documents are more closely related to these activities than may have first been apparent. In fact, it may be the case that exemplary documents actually arise out of the activities that use or generate sets of documents, that is, they may actually play a role in the conduct of the activity. As Xerox Vice-President Steven Kiser put it, "We used to think of a document as the end result of a process; the reality is that the document *is* the process." (van Kirk, 1992) This is clearly the situation in litigation support, where the complaint – which we identified as an exemplary document – is not just an intellectual road map of the lawsuit, but specifies the goals of the lawsuit, or how the case will be litigated. Likewise, in the area of public policy debate, an exemplary document not only can show how the literature can be organized, but it may also describe and perhaps even influence the course of the debate itself. Even a survey of an academic field can be a natural consequence of the activity of that field. That is, academic fields are not just collections of topical information or research. They are ongoing, active research disciplines that are engaged in furthering research in a particular area. An important component of the activity of research in any area is for researchers to be informed about what has been done and what is being done in their area of research. Survey articles specifically address this need for researchers to view the work in their field, so they are, in effect, actually participants in the ongoing activity of research. It is no surprise, then, that they can be good instruments for organizing the textual information of a field.

It is clear that if exemplary documents arise out of activities, then they are likely to assume many of the characteristics of the activities that they come from. Activities, fundamentally, have goals or purposeful concerns. They reflect an intent to accomplish something more or less specific. These goals or concerns provide a point of view or context for the activity. That is, they provide a framework in which a participant can judge what is pertinent to an activity and what is not. It is a fact of natural language that it, too, requires a point of view or context, otherwise, it is difficult to interpret unambiguously (for the simple reason that the same words can be used in so many different activities – that is, they can be used in so many different ways). If natural language is needed to represent documents, as we have stressed in this paper, then it stands to reason that for a language of document representation to be unambiguous it must have a clear point of view or context. This point of view or context is usually what is missing from traditional indexing schemes, but it is what exemplary documents can provide. This also helps to explain why simple full-text retrieval systems, which we discussed in the beginning of the article, have such consistently low levels of recall. By using all the content-bearing words of the documents in a collection to represent those documents for retrieval, the inquirer is presented with an enormously wide variety of access terms and combinations (see Blair & Maron (1985) for examples of this phenomenon). But this is only half of the problem: not only is there

an enormously large number of terms that could provide access to documents in the collection, but there is no broad context in which to interpret the meanings of those words – their purpose or use may be unclear. Exemplary documents address both of these problems: they provide a selected vocabulary to represent a set of documents, rather than simply using *all* the words in the documents to represent them; exemplary documents also provide a narrative context for the selected vocabulary that helps the inquirer to understand what certain specific words mean – that is, how they might be used to represent non-exemplary documents that he might be interested in.

8. Exemplary documents: practical advantages

The use of exemplary documents as an intellectual organizing principle has several specific practical advantages for system design:

8.1. Exemplary documents embody both "top–down" and "bottom–up" organizing principles

Information retrieval systems are often based on an implicit inconsistency, namely, that while a given system's intellectual structure (indexing or representation scheme) is usually designed from the bottom–up, the individual searchers often begin their searches from the top–down. By bottom–up design we mean that the system designers construct the intellectual access structure of the system by an examination of the individual documents included in the system. While this structure can be designed either by individuals (indexers) or it can be done automatically, it still begins in the same place – by evaluating individual documents in the same, often random, order in which they are received. The inquirer, on the other hand, often takes what we could call a top–down approach. He/she does not usually begin with a particular document in mind, but with some more-or-less abstract description of the intellectual content he/she is searching for. How well that abstract description of what the inquirer wants matches with the specific indexed descriptions of documents that he would find useful, is problematic. Since the top–down and bottom–up approaches are based on different principles or different points of view, then it is not likely that they will match up very well. The key then, to bridging the inconsistency between top–down and bottom–up approaches is to develop abstract principles of representation that are based on individual documents. This brings together the inquirer's need for an indexing scheme that can represent high level, abstract concepts with the system designer's need to represent individual documents.

Clearly, you cannot develop a useful schema for representing the abstract intellectual content of a set of documents by basing it on randomly selected documents. Most documents are too narrowly focused to be representative of the implicit intellectual structure of the document collection to which they belong. This is why it is important to identify exemplary documents. Exemplary documents provide an abstract view of the intellectual structure of a set of documents that is grounded in actual documents in that set. Specifically, such documents provide the system designer with not only the abstract intellectual content of a set of documents, but also the relationship or structure of those intellectual abstractions as well as the specific vocabulary in which they are discussed.

8.2. Exemplary documents provide a common foundation for both indexing and searching

As mentioned in Section 4.1 (*supra*), inquirers and system designers/indexers may often begin their respective tasks from different starting points. A clear advantage of exemplary documents is that they provide a common foundation for both the document representation *and* the search formulation process. The indexer, or system designer, will look to the exemplary documents for the indexing vocabulary and structure of the documents he/she wants to represent, while the inquirer will also go to them for the concepts being represented in the collection and the specific vocabulary used to discuss these concepts.

Although there still may be some divergence between the indexer's and inquirer's conception of how the documents might be represented, there appears to us that there will be less divergence if they both begin their tasks in the same place (the importance of this divergence was discussed in Blair (1986)).

8.3. Exemplary documents enable inquirers to better understand the intellectual structure of a document collection

One of the trade-offs we mentioned earlier in this paper was that the more complex the intellectual structure of an indexing or representation scheme is, the harder it may be for the inquirer to understand. Yet, if we want to provide high levels of retrieval effectiveness for some information retrieval systems, then we will need fairly complex and detailed intellectual structures for representing their documents. But such complex representation schemes will require that inquirers must be able to learn them. Unfortunately, most indexed collections of documents are not constructed to facilitate the "learning" of the intellectual structure (Blair, 1990). Often the only way an inquirer can learn how documents are represented on a system is by trial and error, by submitting queries and retrieving documents (Swanson, 1977). Such a piecemeal process is a very unsystematic and probably flawed way to learn, in any general way, how the documents in the collection have been represented. In the first place, the inquirer may never be able to see enough documents to form any general opinion about how the documents are represented (Blair, 1980); and, in the second place, if the inquirer can only see individual documents it may be hard to infer what the broad intellectual relationships are that may exist among groups of documents. Exemplary documents can help to defeat, or at least mitigate, this problem by providing access to not only the intellectual concepts of a field or sub-field, but also to the structure of those concepts and the specific vocabulary used to discuss them.

If an inquirer wants to understand what the intellectual structure of a document collection is, he should begin his search by reading the most relevant exemplary document that exists on the system. The search process, then, is two-tiered: the inquirer searches first through only the exemplary documents available on the system, he then constructs further search queries based on the parts of the exemplary documents that he has found relevant to his search. (In a hyper text-like document management system such as the World Wide Web, links to other, non-exemplary, documents could be embedded in the exemplary documents. In this way, the inquirer would not have to formulate his/her own search queries at every stage in the search.) An exemplary document gives the inquirer a view of the topics or issues germane to her search, a framework that relates these topics or issues, and a clear vocabulary which discusses these topics and can be used

to formulate search requests. (The search procedure based on exemplary documents is similar to the strategy of "seed searching" proposed by Blair (1986).)

8.4. Exemplary documents help to limit the "productivity" of natural language

One of the major characteristics of natural language is its productivity. Language is productive in the sense that a relatively small vocabulary and a few rules of syntax and semantics can combine to produce an uncountably large number of acceptable sentences or descriptions (this is the notion of a "generative grammar"). The productivity of language has proved to be both a virtue and an obstacle for information retrieval systems (as it has for any other computerized system that has a large semantic component): productivity is a virtue because it is relatively simple to find a reasonable way to represent any document; but this productivity is an obstacle, too, because language is so creative that the number of "reasonable" representations for a document and its content is virtually unlimited (Blair, 1990; Swanson, 1966). The same problem holds for the inquirer, too. The inquirer can easily formulate more semantically similar search requests than he/she can possibly have time to actually use, and, unless the inquirer uses them all he/she will have no idea which ones are the most successful.

Since natural language is also productive, as we said, why don't we have the same problem in our everyday discourse as we have in representing documents – that is, while there are lots of semantically equivalent ways to say something, we usually do not have a great deal of trouble expressing ourselves clearly and understanding others. The obvious answer to this is that in everyday speech we have a lot of help from the pragmatic context of our utterances. This is how we can say something quite precise with a sentence that, taken "out of context", would be impossibly ambiguous (see Blair, 1990 for a discussion of this ambiguity problem). For example, suppose that when I arrive at my office I ask a colleague whether so-and-so is at work today. She answers, "I saw a yellow Volkswagen in the parking lot this morning". Strictly speaking, this is not an answer to my question, yet with some simple assumptions about the context in which the statement is uttered and the intentions of my colleague, I can "make sense" of what might be, taken literally, an impossibly ambiguous statement. Such implicatures are an important part of the pragmatic context of speech (Grice, 1989; discussed in Blair, 1992).

> Exemplary documents limit the productivity of language by providing a context of linguistic usage for potential indexing terms. That is, of all the many ways a given word might be used to represent the intellectual content of documents, the exemplary document provides an example of what precise concepts an index term might be used to represent in that collection of documents.

9. Exemplary documents: theoretical justification

While we feel that the practical rationale for the use of exemplary documents in information retrieval is compelling, we also claim that there are sound theoretical reasons that underscore the importance of this kind of document. The theoretical justification for the utility of exemplary

documents is based on the fundamentally linguistic nature of information retrieval. The justification goes like this:

> The effectiveness of an information retrieval system is crucially dependent on how the documents in the collection are represented (for example: assigned keywords; full text; automatically generated index terms; etc.). These document representations are fundamentally linguistic in nature, and are drawn from a subset of natural language discourse. But the meanings of keywords or terms extracted from the text of documents are often not exactly equivalent to the meanings of those same words when used in everyday discourse. For example, the document content that is indexed by the term "computers" is similar to, but may (or may not) be equivalent to "what we mean" when we use the word computers in everyday speech or writing. The word computers, when used to represent the intellectual content of a document, is open to a wide variety of interpretations (that is, computers can be used to represent a wide variety of different, though often related, document topics) (Blair, 1986). Of course, it is also true that the word computers, when used in everyday language and considered by itself, can be quite ambiguous, too. But everyday language has one major advantage over document representation: the words that we use in everyday language have a pragmatic context and accompanying implicatures that can be used to disambiguate individual word uses (Blair, 1992; Grice, 1989). Such a pragmatic context has no equivalent in computerized information retrieval. But for an inquirer to understand what an indexing term means he/she must come to understand the "language of document representation" – and he/she must do so in the absence of any easily accessible contextual information about how the words of that "language" are used.

Putting aside, for the moment, the problem of the lack of contextual reference for the language of document representation, let us look briefly at how a mature individual learns natural language (here we will assume that we are dealing with an intelligent adult who is either learning a new language entirely, or learning new words in an already familiar language). It is no great intellectual leap to see that it would be best for the inquirer to learn the language of document representation in the same way that he would learn new words in any natural language. How is this done? While there are many theories of language acquisition, the theory proposed here is based on the later philosophy of language of Wittgenstein. One of the present authors has argued that Wittgenstein's later philosophy of language is of significant use in understanding the unique problems of document representation (Blair, 1990).

10. Language acquisition and übersichtliche Darstellungen

Briefly, Wittgenstein's theory of language acquisition runs like this: We do not acquire language by definitions and explanations alone, but by having the *use* of the expressions in question demonstrated to us – ideally, in the same circumstances or activities in which they are ordinarily used. Definitions and explanations are not central to language acquisition and can be entirely replaced by examples and demonstrations of word usage. Definitions and explanations *do* have a use in language acquisition, but it is not a central use. They serve to clarify or modify a particular

linguistic expression whose general type of usage is already relatively clear. The examples of language usage which are used in training must be "perspicuous" – they must be clear and compelling. In short, Wittgenstein's theory of language acquisition is crucially dependent on the identification of "perspicuous representations" of word usage. [1]

It stands to reason, that if Wittgenstein's philosophy of language acquisition is correct, or at least plausible, then to enable inquirers to learn how document representations are used we must provide them with examples of not just *any* index term usage, but with perspicuous term usage. These representations of usage must be exemplary, clear, and embedded in a rich enough linguistic context that their usage is relatively unambiguous. Within the linguistic framework of information retrieval, the usage and context of index terms is not at all clear, in fact, as we mentioned before, the pragmatic context of index term usage is virtually non-existent. (Recalling our previous example, it would be like trying to discover the meaning of my colleague's statement "I saw a yellow Volkswagen in the parking lot this morning" by just tabulating the meanings of the individual words in the sentence without regard to the circumstances in which it was uttered.) Further, if one were to look at actual index term assignments to documents it would be difficult to distinguish the term assignments that are exemplary from those that are not. This is why exemplary documents are so important. They not only provide a view of a topic area (in whole or in part), but they also are an active demonstration of the language that is used to talk about these areas – in short, exemplary documents provide the perspicuous representations of index term usage that would permit the inquirer to understand what these terms might mean if they were used to represent the intellectual content of other documents. The text of exemplary documents also provides the pragmatic context of index term usage that is missing in contemporary information retrieval systems.

11. Exemplary documents: implementation

As mentioned earlier in this paper, we are currently using the notion of exemplary documents as a basis for the design of information retrieval systems. It might be useful to sketch some of the

[1] What has been translated as perspicuous representations is what Wittgenstein originally called "übersichtliche Darstellungen". This is not an unequivocal translation for there is no simple English equivalent (see Baker and Hacker (1985) for a discussion of the translation problem here), but the essential concepts which constitute übersichtliche Darstellungen are that they are not just perspicuous examples of word usage, but representations that give us a "broad view" of how the word in question is used. That is, any representation of a word usage is not necessarily a perspicuous one. The representation must be, to use another word, exemplary. In addition, the German word "Darstellung" means not just "representation", but also "performance". The notion of a performance is crucial to Wittgenstein's philosophy of language acquisition – since he emphasized the importance of how words are used in the conduct of our daily activites and practices, rather than how they are defined. In short, what the individual learning new words or a new language needs is a good supply of exemplary uses or performances of those words within the context of natural language discourse.

> A main source of our failure to understand is that we do not *command a clear view* of the use of our words. Our grammar is lacking in this sort of perspicuity [Übersichtlichkeit]. A perspicuous representation [Darstellung] produces just that understanding which consists in 'seeing connexions'. (N.B."connexions" is the translator's rendering of "Zusammenhänge" which can also be translated as "context") (Wittgenstein, 1953, para 122).

directions we are taking. Since exemplary documents actually embody the intellectual structure of the literature they are found in, it may be useful to implement at least part of this organization through some kind of hypertext system. In that way, links to non-exemplary documents could be embedded right in the text of the exemplary documents, or could be generated automatically based on the similarity between the text of the exemplary document and the representations of the non-exemplary documents. The advantage of this is clear: when inquirers read a section of the exemplary document that they are interested in, they could activate such a link and move directly to those documents most closely related to the section they were reading. This would save inquirers from having to formulate a search query for each section of an exemplary document that they are interested in. If a number of documents were identified by the same link, then they would, of course, need to be ranked by the closeness of fit between them and the section of the exemplary document to which they are related. There are a number of existing procedures that could be employed to do this.

It is clear that simple hypertext links are just one of the ways in which exemplary documents can be used to organize a collection of documents. Many of the nodes on the WWW are already organized this way. An improvement might be to write or find exemplary documents that could be used as the first level of access at the home page of a web site.

12. A final note on coverage

One important aspect of exemplary documents that may trouble some readers is that there may be cases where exemplary documents exist, but do not provide complete coverage of all the intellectual issues of a field or sub-field. It may also be the case that they might not provide the granularity of description desired for that field. Until a concerted effort is made to identify exemplary documents in a field, it is uncertain how complete the intellectual road map they provide will be. It is hoped that the coverage provided by exemplary documents will be adequate, but short of that, the coverage that they do provide, and the context of representation that they bring with them, will be useful adjuncts to the traditional manual or automatic methods of document representations that we currently employ.

13. Conclusion

This article has addressed the specific problem of constructing document representation schemes in situations where traditional representation schemes do not apply well, and where highly effective access to the documents is required. We have proposed that certain documents, which we call exemplary, may exhibit useful ways for representing a class of documents. These exemplary documents provide several advantageous things: an intellectual road map to the topics and issues of a field; a specific vocabulary that can be used to refer to those topics and issues; and an interpretive context in which the language used to represent a set of documents can be disambiguated. We have also argued that apart from several distinct practical advantages for the use of exemplary documents, there are also several theoretical reasons that also support the use of exemplary documents in information retrieval situations.

Acknowledgements

The authors wish to thank Pat Wilson and M.E. Maron (University of California, Berkeley) for their comments on an earlier version of this paper.

References

Austin, J. (1962). *How to do things with words*. Oxford: Oxford University Press.

Baker, G. P. & Hacker, P. M. S. (1985). Ubersicht. In *Wittgenstein: Meaning and understanding. Essays on the philosophical investigations. Vol. 1*. The University of Chicago Press; p. 295–310.

Blair, D. C. (1992). Information retrieval and the philosophy of language. *Computer Journal, 35*(3), 200–207.

Blair, D. C. (1990). *Language and representation in information retrieval*. Amsterdam: Elsevier.

Blair, D. C. (1986). Indeterminacy in the subject access to documents. *Information and Processing Management, 22*(2), 229–241.

Blair, D. C. (1980). Searching biases in large interactive document retrieval systems. *Journal for the American Society for Information Science, 31*(4), 271–277.

Blair, D. C., & Maron, M. E. (1985). An evaluation of retrieval effectiveness for a full-text document retrieval system. *Communications of the ACM, 28*(3), 289–299.

Blair, D. C., & Maron, M. E. (1990). Full-text information retrieval: further analysis and clarification. *Information Processing and Management, 26*(3), 437–447.

Bruner, J. (1990). *Acts of meaning*. Cambridge, MA: Harvard University Press.

Bush, V. (1990). As we may think. *The Atlantic Monthly, 176*, 101–108.

Delphi Consulting Group Inc. (1993). *Text retrieval systems: A market and technology assessment*. Boston, MA.

Clinton/Gore '92 Committee. (1992). *Technology: The engine of economic growth. A national technology policy for America*. National Campaign Headquarters, Little Rock, Arkansas. Final Version, September 21.

Fleischer, R. (1991). Total document control: a text-retrieval perspective. In G. Peter (Ed.), *Text retrieval: information first, Proceedings of the Institute of Information Scientists 1990 Text Retrieval Conference*, London, October 1990. London: Taylor Graham.

Grice, H. P. (1989). *Studies in the way of words*. Cambridge, MA: Harvard University Press.

Hilbert, D. (1901). Mathematische probleme. *Archiv für Mathematik und Physik*. 3 Reihe Bd. 1, S. 44–63; S. 213–237.

Kimbrough, S. O., Clark, P., Michael, B., & Hemant, B. (1990). The coast guard's KSS project. *Interfaces, 20*(6), 5–16.

van Kirk, D. (1992). Document management: Destined to become a smash hit. *Infoworld*, November 2, p. 52.

Krishnan, R. (1993). Model management: survey, future directions and a bibliography. *ORSA CSTS Newsletter, vol. 14(1)*, Spring.

Kuhn, T. S. (1970). *The structure of scientific revolutions* (2nd ed.). Chicago, IL: University of Chicago Press.

Langendoen, D., & Postal, T. P. (1984). *The vastness of natural languages*. Oxford, UK: Basil Blackwell.

National Science Board. (1992). *The competitive strength of US industrial science and technology: Strategic issues*. Committee on Industrial Support for R&D, August.

Salton, G. (1986). Another look at automatic text-retrieval systems. *Communications of the ACM, 29*(7), 648–656.

Searle, J. (1969). *Speech acts: An essay in the philosophy of language*. Cambridge: Cambridge University Press.

Swanson, D. R. (1977). Information retrieval as a trial-and-error process. *Library Quarterly, 47*(2), 128–148.

Swanson, D. R. (1966). *Studies of indexing depth and retrieval effectiveness*. Unpublished report, National Science Foundation Grant GN 380, February.

Turner, M. (1996). *The literary mind*. Oxford: Oxford University Press.

Will, H. J. (1975). Model management systems. In E. Grochia & N. Szyperski (Eds.), *Information Systems and Organization Structure*, (pp. 468–482). Berlin: Walter de Gruyter.

Wittgenstein, L. (1953). *Philosophical investigations* (G.E.M. Anscombe, Trans.). New York: The MacMillan Company.

Available online at www.sciencedirect.com
SCIENCE DIRECT
Fuzzy Sets and Systems 148 (2004) 5–19
FUZZY sets and systems
www.elsevier.com/locate/fss

Web mining: a survey in the fuzzy framework☆

Dragos Arotaritei[a], Sushmita Mitra[b,*]

[a] *Department of Computer Science, Aalborg University Esbjerg, Niels Bohrs Vej 8, 6700 Esbjerg, Denmark*
[b] *Machine Intelligence Unit, Indian Statistical Institute, 203 B. T. Road, Kolkata 700108, India*

Abstract

This article provides a survey of the available literature on fuzzy Web mining. The different aspects of Web mining, like clustering, association rule mining, navigation, personalization, Semantic Web, information retrieval, text and image mining are considered under the existing taxonomy. The role of fuzzy sets in handling the different types of uncertainties/impreciseness is highlighted. A hybridization of fuzzy sets with genetic algorithms (another soft computing tool) is described for information retrieval. An extensive bibliography is also included.
© 2004 Elsevier B.V. All rights reserved.

Keywords: Fuzzy sets; Soft computing; Web mining; Information retrieval

1. Introduction

Web mining refers to the use of data mining techniques to automatically retrieve, extract and evaluate (generalize/analyze) information for knowledge discovery from Web documents and services. Web data is typically unlabelled, distributed, heterogeneous, semi-structured, time varying, and high dimensional. Hence any human interface needs to handle context sensitive and imprecise queries, and provide for summarization, deduction, personalization and learning. Almost 90% of the data is useless, and often does not represent any relevant information that the user is looking for. Taking into account the huge amount of data storage and manipulation needed for (say) a simple query, the processing essentially requires adequate tools suitable for extracting only the relevant, sometimes hidden, knowledge as the final result of the problem under consideration.

The use of soft computing tools, including fuzzy logic, in data mining has been adequately reported in literature [24]. However, the subtle differences between data mining and Web mining suggests the

☆ This work was carried out when Dr. S. Mitra was a Visiting Professor at the Department of Computer Science, Aalborg University Esbjerg, Denmark.

* Corresponding author.
E-mail address: sushmita@isical.ac.in (S. Mitra).

0165-0114/$ - see front matter © 2004 Elsevier B.V. All rights reserved.
doi:10.1016/j.fss.2004.03.003

use of new or modified tools and algorithms for appropriate handling of the Internet. Web mining typically addresses semistructured or unstructured data, like Web and log files with mixed knowledge involving multimedia, flow data, etc., often represented by imprecise or incomplete information. This implies that fuzzy set theoretic approaches are useful instruments in order to mine knowledge from such data.

The huge amounts of multivariate information offered by the Web has opened up new possibilities for many areas of research. Due to the involvement of human interaction in Web information (like text, image, sound, and linkages between them), new tools and methodologies need to be extended in order to deal with the incomplete or imprecise information. Web personalization, navigational Web, semistructured and structured data information are some of the major issues under Web mining. Semantic Web is also a target for data mining research using fuzzy logic, because of the inherent vagueness of human beings when expressing information in natural language. Considering the Web as a large distributed multimedia database, an extension of methodologies to deal with them and their mining algorithms can be considered under image mining. The huge volumes of compressed and uncompressed information stored in images, and their subjective evaluation by humans in interaction with the Web, is another focus of current data mining research requiring fuzzy treatment.

Web mining can be broadly categorized as

- Web Content Mining of multimedia documents, involving text, hypertext, images, audio and video information. This deals with the extraction of concept hierarchies/relations from the Web, and their automatic categorization.
- Web Structure Mining of inter-document links, provided as a graph of links in a site or between sites. For example, in Google a page is important if important pages point to it.
- Web Usage Mining of the data generated by the users' interactions with the Web, typically represented as Web server access logs, user profiles, user queries and mouse-clicks. This includes trend analysis (of the Web dynamics information), and Web access association/sequential pattern analysis.

Fig. 1 provides a taxonomy for Web mining. The different functions falling under the three major categories are highlighted. A relationship with information retrieval is also depicted.

The role of fuzzy sets in Web mining [5] holds promise mainly in (i) document and user clustering, (ii) deduction and summarization, (iii) handling of fuzzy queries involving natural language and/or linguistic quantifiers like *almost*, *about*, and (iv) information fusion in multimedia data. According to Zadeh [35], fuzzy logic may serve as the backbone of the Semantic Web, an extension of the current Web in which information is given well-defined meaning, thereby better enabling computers and people to work in cooperation.

Ordinary end-users often face difficulties in formulating a precise representation of their information needs in a Boolean query. This affects the efficiency of the information retrieval process. Hence Web search engines require the use of fuzzy aggregation operators. These are specially suitable in flexible query answering and information retrieval. User feedback is a process where the user expresses her/his opinion about the documents that the system has retrieved as an answer to a certain query. This user evaluation/opinion can be used both for training the classification system as well as reflected in the corresponding user profile. Fuzzy linguistic modeling is useful in handling such preference relations. This information can also be utilized by (say) marketing experts to analyze user interests.

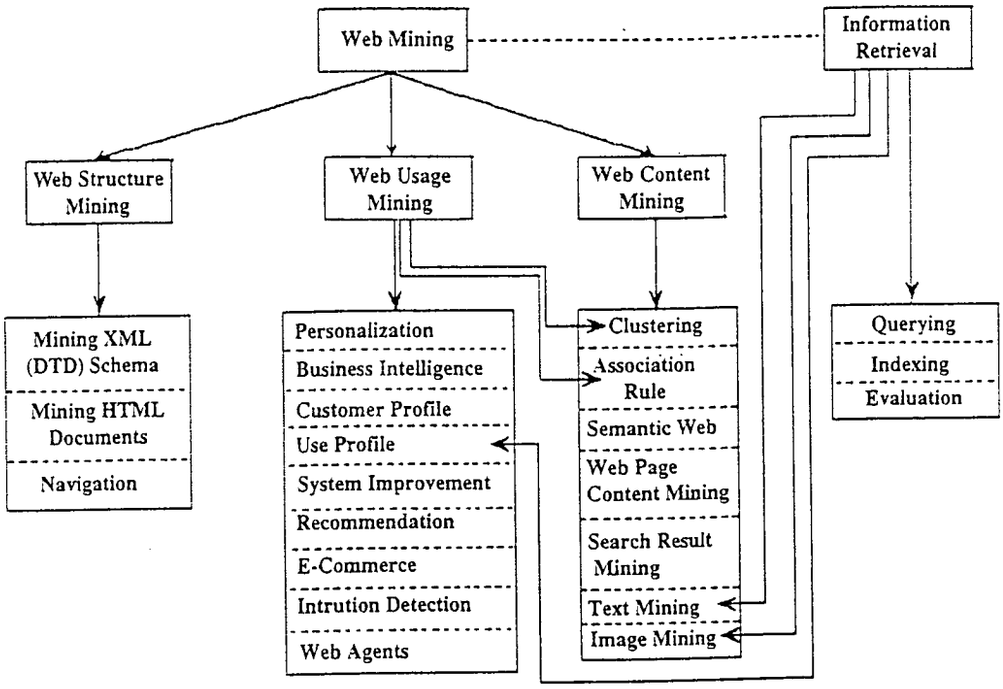


Fig. 1. A web mining taxonomy.

In this article we present a survey on the use of fuzzy sets in the different aspects of Web mining. In Sections 2–9 we dwell on varied issues like Web clustering, association rule mining, Web navigation, Web personalization, Semantic Web, information retrieval, text mining and image mining. The article is concluded in Section 10.

2. Web clustering

Clustering pertains to unsupervised learning, where no predefined classes are assigned. The key requirement is the need for a good measure of similarity between the instances/patterns. The problem is to group n patterns into c desired clusters, such that the data points within clusters are more similar than across clusters. Scalable clustering algorithms pertain to working with large volumes of high-dimensional data, that is inherent to data mining problems. Moreover, the presence of mixed media data over the Web call for the use of specialized techniques.

The use of fuzzy logic has been widely studied in clustering problems [1], and the results have also been extended to handle high dimensional data [17,28]. The importance of clustering to Web mining, specifically in the domains of Web Content and Web Usage mining, make Web clustering an interesting topic of research. This includes clustering of Web documents, snippets and access logs. Usually the Web involves overlapping clusters. So a crisp usage of metrics is better replaced by fuzzy sets which can reflect, in a more natural manner, the degree of belongingness/membership to a cluster.

In the fuzzy c-means algorithm, the objective is to minimize the function

$$J(U,V)=\sum_{i=1}^{c}\sum_{j=1}^{n}u_{ij}^{m}diss(v_i,x_j) \quad \text{subject to} \quad \sum_{i=1}^{c}u_{ij}=1 \text{ for all } j, \tag{1}$$

where $diss(v_i,x_j)$ is the distance of pattern x_k from the ith cluster mean (typically expressed as squared Euclidean distance), u_{ik} is the membership function (in the interval [0,1]) of point x_k in the ith cluster such that $0<\sum_{j=1}^{n}u_{ij}<n\ \forall i$, U is the fuzzy c-partition of the data set, V is a set of c-prototypes and $m>1$ is the fuzzifier. However, even a few outliers or inherent noise in real data can affect the result of this algorithm. Fuzzy c-means has problems finding correct clusters in the presence of noise or outliers, because of its assumption that any point in a dataset must essentially belong to a cluster.

Using a noisy cluster concept in [6], all the noise points are collected in a separate cluster. These noise points are supposed to have a distance $diss(v_c,x_k)=\delta$ from their prototype v_c. Any pre-specification of δ is not easy, and the distance can be different for different kinds of problems. One way is to use a simplified statistical average to compute δ as

$$\delta=\lambda*\left[\frac{\sum_{i=1}^{c}\sum_{k=1}^{n}diss(v_i,x_k)}{n(c-1)}\right], \tag{2}$$

where λ is a factor experimentally determined to lie between 0.1 and 100.

The probabilistic constraint that the membership u_{ij} of a data point x_j have a sum of one, is relaxed in the possibilistic approach to fuzzy c-means [18]. Here we have

$$\max_i\{u_{ij}\}>0 \quad \text{for all } j,$$

in Eq. (1). In order to avoid the trivial solution $u_{ik}=0$, a penalty function for low membership is appended to the objective of Eq. (1). The objective function now becomes [18]

$$J_P(U,V)=\sum_{i=1}^{c}\sum_{j=1}^{n}u_{ij}^{m}diss(x_j,v_i)+\sum_{i=1}^{c}\alpha_i\sum_{j=1}^{n}(1-u_{ij})^m, \tag{3}$$

where the α_i's are a set of chosen numbers. The first part of this equation minimizes the distance of the feature vectors from the prototypes, while the second term maximizes the membership degree of this vector to the corresponding prototype. The membership functions u_{ij} are viewed as possibility distributions for clusters U_i over the domain of discourse of feature vectors x_j. The terms α_i are chosen based on the shape and size of cluster U_i, and the relative weight assigned to the second term in Eq. (3). A typical choice is to make α_i proportional to the average fuzzy intra-cluster distance of cluster U_i. Depending on the distance measure $diss(x_j,v_i)$ used, one derives the possibilistic c-means, possibilistic Gustafson–Kessel and possibilistic c-quadric shells algorithms.

A generalized model called alternating cluster estimation (ACE) [30] uses alternating iterations on model architectures, while the membership and prototype functions are selected directly by the user. (S)he can choose the form of updating of equations that calculate the new partition $U(t)$ and new prototypes $V(t)$ at each iteration t. The pair of updating equations can be changed in order to produce a model that matches better with the data to be clustered. Selection of membership function families (triangular, hyperconic, etc.) by approximating from available data, and the prototype function builder are the basics of an ACE interface.

A review of robust methods used for fuzzy Web clustering is presented in [13]. A robust algorithm is one whose performance is minimally affected in the presence of outliers. The authors present the fuzzy c-means (FCM), fuzzy trimmed c-prototype, fuzzy c-least medians and relational fuzzy c-means algorithms, along with some techniques used to make them robust.

The fuzzy c-medoids (FCMdd) algorithm can be summarized as follows. Let $X_c = \{x_i \mid i = 1, \ldots, n\}$ be a set of objects, $diss(x_i, x_j)$ be the matrix of dissimilarity between the objects x_i and x_j and $V = \{v_i \mid i = 1, \ldots, c\} \subseteq X_c$ be the set of representative points. The objective function E_m that must be minimized over all V in X_c is given as

$$E_m(V, X_c) = \sum_{i=1}^{c} \sum_{j=1}^{n} u_{ij}^m diss(x_j, v_i), \tag{4}$$

where u_{ij}^m represents the fuzzy or possibilistic membership of x_j to cluster i with a heuristic determination of m (that controls the shape of this membership function). The steps of the algorithm are as follows:

(i) Set the number of clusters c
(ii) Randomly pick up initial set of medoids V from X_c
(iii) Set $iter = 0$
(iv) For $i \leftarrow 1$ to c do // Compute the membership
  For $j \leftarrow 1$ to n do
   Compute u_{ij}^m, using eqn. (6), for $x_j \neq v_p$
  endfor
 endfor
 $V^{old} = V$ // Store the current medoids
 // Compute the new medoids
 For $i \leftarrow 1$ to c do

$$q = arg \min_{1 \leqslant p \leqslant n} \sum_{j=1}^{n} u_{ij}^m * diss(x_p, x_j) \tag{5}$$

$$v_i = v_q$$

(v) If $iter = MAX_ITERATIONS$ or $V^{old} = V$, then stop else go to step 4

The membership function is expressed as

$$u_{ij}^m = \frac{\left(\dfrac{1}{diss(x_j, v_i)}\right)^{1/(m-1)}}{\sum_{p=1}^{c} \left(\dfrac{1}{diss(x_j, v_p)}\right)^{1/(m-1)}}. \tag{6}$$

For $m = 2$ this boils down to the harmonic mean of the dissimilarities. The sum of square criterion is used in the dissimilarity matrix, typically expressed as the distance in the FCM and FCMdd algorithms. This is found to be not very robust.

In the robust FCM, one way of achieving robustness is by modifying Eq. (4) to

$$E_m^{\mathrm{R}}(V,X_c)=\sum_{j=1}^{n} g\left(\left(\sum_{i=1}^{c}(diss(v_i,x_j))^{1/(1-m)}\right)^{1-m}\right), \tag{7}$$

where g is a loss function that reduces the role of outliers and v_i refers to the ith class mean. Other approaches include possibilistic c-means [18], noise clustering [6] and alternating cluster estimation [30], as discussed earlier in this section.

The fuzzy c-medoids and fuzzy c-trimmed medoids are used to cluster relational data from Web documents and snippets in [16]. The algorithms are applied to a collection of 1042 abstracts from the Cambridge Scientific Abstract Web site, corresponding to 10 topics. A preprocessing stage is used to filter and remove irrelevant words, in order to generate the input feature vector that is computed using an *inverted document frequency* method. This 500-dimensional feature vector (keywords) is reduced using principal component analysis, resulting in a selection of 10 eigenvector values. The algorithms are also tested on a collection of snippets, corresponding to 200 Web documents collected by a search engine in response to a query "salsa". The results are extended in [17] using robust estimators, providing a computational complexity $O(n \log n)$ instead of $O(n^2)$. This is suitable for clustering noisy data that are characteristic of Web documents. An application of the robust clustering algorithm is made for mining Web access logs in [14], in order to automatically discover user session profiles.

The ACE model for numerical values [30] is extended to handle non-numerical patterns in relational data sets using relational ACE (RACE) [29]. The authors consider two types of patterns related to the Web, viz., (i) document contents such as text parts of Web pages (Web content mining), and (ii) sequences of Web pages visited by users, such as Web logs (Web usage mining). Both text and Web page sequences are handled using the Levenshtein (edit) distance, that determines the number of edits (delete, insert, change) steps needed to convert one word into another, and a graph-based distance measure. The prototypes found for Web page sequences are interpreted as typical click streams reflecting prototypical user interests. This provides an automatic data-driven user segmentation of the Web page sequences.

3. Association rule mining

Association rules, in the context of Web mining, refer to the determination of (say) those URLs that tend to be requested together. This can be categorized under Web Usage and Web Content mining. In this section we describe some methodologies that incorporate fuzzy set theoretic concepts.

Web mining of inference amplification is made in [21], using inference logic from fuzzy cognitive maps in three phases. The first stage mines association rules with the a priori algorithm, from a Web-log database. A corresponding fuzzy knowledge map involving causality (the cognitive map uses values between -1 and $+1$) is built in the second stage, incorporating both positive and negative rules. In the final stage, the system applies inference amplification in order to enrich the resulting Web mining association rules. Experiments are made using a Web-log file from a real online shop, specializing in computers and peripherals. The causal knowledge is represented as an adjacency

matrix, including the connectivity $W^{\mathrm{T}} = \{w_{ij} \mid w_{ij} \in \{-1, +1\}\}$. Here w_{ij} indicates the causal value (weight) of the arc from vertex/node C_i to C_j.

The definition "A causally increases B", implies that an increase/decrease of A causally increases/decreases B. The equivalent fuzzy relation (causal knowledge) used for inference amplification is expressed as

$$C_i \overset{w_{ij}}{\rightarrow} C_j;\ C_i \overset{\sim w_{ij}}{\rightarrow} \sim C_j;\ \sim C_i \overset{\sim w_{ij}}{\rightarrow} C_j;\ \text{and}\ \sim C_i \overset{w_{ij}}{\rightarrow} \sim C_j,$$

where $\sim$ indicates negation, and w_{ij} is the fuzzy membership function value $R(C_i, C_j)$ in the fuzzy relation between the fuzzy concept nodes C_i and C_j. After the chained rule is extracted, the inference amplification is made according to the above rules. The basic case $T \rightarrow T$, with $T = \textit{true}$ and $F = \textit{false}$, in the rule

$$A\ 0.2\ (T) \rightarrow 1.0\ B\ (T) \rightarrow 0.84\ C\ (T)$$

signifies that 20% of A browse B, and that 84% of them buy the product C. When inference amplification is applied, it produces the additional subtle information

$$A\ 0.8\ (F) \rightarrow 1.0\ B\ (T) \rightarrow 0.16\ C\ (F).$$

Here the case $F \rightarrow F$ is extended in a manner analogous to the case $T \rightarrow T$.

Fuzzy association rules are mined to discover access path prediction in [34]. The authors use a fuzzy index tree, involving fuzzy cases, in order to reduce the explosion in the number of rules generated in the corresponding crisp state. The experiments use the data from the logs made in www.microsoft.com.

A meta-search engine based on query keywords, that uses instant information retrieval given by Grey relational method, is described in [12]. A fuzzy inference model is used in the common case when the number of keywords is greater than or equal to two. The related sites are mined in order to verify the users' expectation. The term frequencies and document frequencies are fuzzified according to nine given rules. The meta-search engine finally locates the closest related site, based on keywords, from the queries.

4. Web navigation

Navigation is categorized as Web Structure mining. An optimization of the path for surfing the Web, given a target, is described in [4]. The connected Web sites are represented by a directed graph with source and destination nodes, and a links set along the path. The frequency of accessing various links (access rate) and the time taken to retrieve target pages (retrieval rate) are considered as the decisive factors. These are affected by criteria like the availability of channels, server capability and access time. The expected access rate and the required retrieval rate are expressed as fuzzy sets. Link weights are associated with appropriate linguistic values. A fuzzy opinion matrix expresses subjective estimation of users during their surfing of specified links on the path. The general evaluation of the link, with respect to interest rate of users, is provided as the intersection among the estimation made by the different users. The optimal path is computed as the minimum fuzzy distance estimated between a fuzzy *Hurwicz opinion set* (derived in terms of optimism–pessimism index) and the actual requirement set.

Web mining is applied to log files in order to discover user's behavior [11]. Numerical data, representing the time spent by an user in a page, is quantified using linguistic membership functions. The domain of discourse [0,130] is split into three trapezoidal membership functions, viz., short, middle and long. The preprocessing phase consists of sorting the log files by client and time access, and mapping the information to a flat database that is mined as multiple items transaction in the fuzzy framework. Maximally large sequences are discovered as a list of most frequent successively traversed Web pages, that satisfy the minimum support criterion for fuzzy association rules. However the fuzzy rule mining algorithm used to identify the sequential navigational patterns is complex, and costly in terms of time and memory requirements if the number of users and number of pages is high (say, thousands of pages). As a consequence, it is possible to have very long frequent sequences comprising of hundreds of items.

5. Web personalization

The increasing popularity of the Internet and the exponential increase in the number of its users has led to the creation of new paradigms of knowledge discovery, like Web personalization, mining bookmarks, mining e-mail correspondences, recommendation systems, and so on. These are grouped as Web Usage mining.

Mining typical user profiles and URL associations from the vast amount of access logs is an important component of Web personalization, that deals with tailoring a user's interaction with the Web information space based on information about her/him. Nasraoui et al. [26] have defined a *user session* as a temporally compact sequence of Web accesses by a user and used a dissimilarity measure between two Web sessions to capture the organization of a Web site. Automatic discovery of user session profile is made using fuzzy competitive agglomeration for relational data (CARD) algorithm. Complex, non-Euclidean distance/similarity measures can be handled in this framework. The objective function is now defined as

$$E^A(V,X_c) = \sum_{i=1}^{c}\sum_{j=1}^{n} u_{ij}^2 \cdot g\left(diss(x_j, v_i)^2 - \alpha \sum_{i=1}^{c}\left(\sum_{j=1}^{n} w_{ij}.u_{ij}\right)^2\right) \tag{8}$$

subject to $\sum_{i=1}^{c} u_{ij} = 1$ for all j, where α is a function of the iteration number, $g(.)$ is the loss function and $w_{ij} \in [0,1]$ are the robust weights. Log files are collected from a real site in order to mine the user profiles. The experimental results "capture" the pattern for users in profiles like outside visitor, prospective student, those that attend lessons by the same professor, an so on. To discover user profiles from real log files, various fuzzy clustering methods have been applied. The robust fuzzy c-medoids [14] is one such example. Another approach in this direction is the relational alternating cluster estimation [29].

A system that looks for Web documents using link-based search and a fuzzy concept network is described in [15]. The user's subjective interests are appropriately represented by a fuzzy concept network based on user profile. Missing information is inferred from a transitive closure of a matrix of knowledge in the network. The degree of relevance in the network is fuzzified as a value between 0 and 1. The importance of the document is computed by the fuzzy retrieval system that personalizes the results given by the search engine. It ranks the retrieved documents, and finds the authoritative

and hub sources. Five best authoritative sources for query are selected as the most representative documents corresponding to a user's query. Documents are ranked according to the user's interests stored in her/his profile (say, ten concepts as "Book", "Java", etc.) The experimental results present the ranked evaluation by three users, and the personalized result in response to their query for "Java".

A Fuzzy Cognitive Map (FCoM) is used in [23] for a causal modeling of the behavior of the user during her/his search on the Web. The learning of connections in the FCoM is made by neural learning of the weight matrix, using Hebbian law, when a change is detected. FCoM evolution proceeds along Markovian principles, by calculating the next concept based on the current concept and edge values. The convergence to a stable limit cycle or fixed points determines the "hidden patterns" in the causal Web represented by the FCoM. It reflects the users' behavior as their knowledge of the Web increases with time, and they either learn new patterns or reinforce old ones. Unlike Markov models, FCoM is able to detect hidden rules related to certain behavior of users like learning from experience (previous mistakes).

6. Semantic web

Despite the huge amount of effort made to tackle the information from the Web, the area of Semantic Web is yet to be fully exploited. Although fuzzy sets offer a possibility to work with linguistic terms, there does not exist much literature on the fuzzy Semantic Web. A fusion between conceptual graph and fuzzy formalism is made in [3] to handle natural language from hypermedia and hypertext semantics. A fuzzy conceptual graph (FCG) is used for knowledge representation of the Semantic Web, and is available both for human and machine processing. Concepts of simple FCG, fuzzy entity and fuzzy attribute are defined. FCG projection is used for matching graphs, and this helps evaluate the degree of matching of the concept and relation pair between two FCGs. Absolute and relative quantifiers from natural language are represented by a lambda expression. For example, the pages about "Books written by A" and "Books written about A" are indexed, and the result of the query is adequately generated. The results have also been extended for indexing images.

However, developments regarding fuzzy conceptual graphs are at an early stage now. Their extension to very large documents, structured and unstructured types of documents with possibly very long sentences, multimedia Web pages, etc., are subjects of ongoing investigation.

7. Information retrieval

Typically, four main constituents can be identified in the process of information retrieval from the Internet.

- *Indexing*: generation of document representation
- *Querying*: expression of user preferences through natural language or terms connected by logical operators
- *Evaluation*: performance of matching between user query and document representation
- *User profile construction*: storage of terms representing user preferences, specially to enhance the system retrieval in future accesses by the user

Due to the presence of *multimedia* information repositories consisting of mixed media data, the information retrieved can be text as well as image document or a mixture of both. An example for mining multimedia data using fuzzy methods is given in Ref. [27]. We deal with text and image mining, separately, in the following two sections in order to highlight their importance in Web mining.

Fuzzy genes are used as intelligent agents for information retrieval from the Web [22]. This incorporates a hybridization of fuzzy sets with genetic algorithms (GAs) in the soft computing framework. User profiles are built from the user preferences, represented by chromosomes made up of a vector of fuzzy genes. Each chromosome is associated with a fitness corresponding to the system's belief in the hypothesis that the chromosome, as a query, represents the user's information needs. Every gene represents, by a fuzzy set, the number of occurrences that characterizes the documents considered relevant by the user. The fitness of the chromosome is adjusted based on the comparison between the user's evaluation of the retrieved documents and the score computed by the system. GAs are used to track the user's preferences and adapt the profile by incorporating her/his relevance feedback, while fuzzy sets handle the imprecision in the user's preferences and evaluation of the retrieved documents.

The architecture of the system has a parser module that collects words and performs statistics, a learning module that uses GAs for modeling the behavior, and an evaluation module that scores the documents. Chromosomes with fuzzy genes are present in the learning module, and use three types of membership functions in the universe of discourse involving term occurrence values. The classical GA operations like selection, crossover and mutation are applied. The fitness function is given as

$$f_i^j = f_i^{j-1} + P_i^j - L_i^j, \tag{9}$$

with

$$P_i^j = [(1 - (C_i^j - U^j))^2]^{w_1} \cdot [(S^{j-1} - U^j)^2]^{w_2},$$

where f_i^j is the fitness of the ith chromosome in the jth generation, P_i^j corresponds to the payoff (used for accumulation of fitness), L_i^j is the lifetax (used as a restriction on the chromosome if the payoff is continually low), and w_1, w_2 are the weights (lying in the unit interval) assigned by experts. Here C_i^j is the score of the ith chromosome in generation j, U^j represents the user's evaluation of the document and S^j is the population score computed by an arithmetic average of the best chromosomes. $C_i^j = 1/\left(l \cdot \sum_{h=1}^{l} \mu_r^h\right)$, where l represents the length of the chromosome and μ_r^h is the membership value of the hth gene of function type r. The performance is evaluated in terms of a prediction precision of the feedback from the user, recapitulation ability, and the classification predictability (CP) expressed as

$$CP = 1 - \sqrt{\frac{1}{d-1} \sum_{i=1}^{d} (S^i - U^i)^2}$$

with d as the number of documents in the evaluated set. The authors use a set of 20 documents with a manual assignment of scores from 0.1 to 0.9 in the search history. In the test procedure the system receives 65 documents split in 13 sets of 5 documents each.

A system composed of WebAgents and MetaWebAgents is described in [2]. While the WebAgents retrieve certain information, the job of the MetaWebAgents is to select the appropriate WebAgents that can access the necessary information. Two characteristics, viz., *reply-time* (answer time from any WebAgent after a MetaWebAgent requests for information) and the number of solutions, are used for measuring the performance of the WebAgents. A fuzzy distance between the performance values of the WebAgents is computed to classify their behavior. A Fuzzy Relational Algorithm (FRA) in the *If ... And ... Then ...* form of Mamdani-implication is used to store the required knowledge. The defuzzification is made using the Center of Gravity method. This system, involving search and meta-search, is used in real problems related to querying for airlines flights.

An approach in Ref. [20] uses fuzzy association thesaurus and query expansion for information retrieval. Fuzzy composition operations like max–min, max–product and sum–product are used for constructing the thesaurus. Interactive query expansion shows the user, upon initial query, a ranked list of documents suggested by the system based on the fuzzy relation composition. A measure of similarity helps select the correlated terms using queries with/without weight. The experiments use a collection of daily news in Chinese, with 981 homogeneous and 700 heterogeneous text documents.

An HTML document can be viewed as a structured entity, in which document subparts are identified by tags and each such subpart consists of text delimited by a distinct tag. A fuzzy representation of HTML documents is described in Ref. [25]. The HTML document is represented by a sum of fuzzy set terms

$$R(doc) = \sum_{t \in T} \frac{\mu_{doc}(t)}{t}, \tag{10}$$

where the importance of each term t in document doc is given by the membership value $\mu_{doc}(t) = \sum_{i=1}^{n} \left(\mu_{tag_i}(doc, t) * w_i\right) * g(IDF_t)$, w_i is the normalized importance weight associated with tag_i, $g(.)$ is a normalization function and IDF_t is the inverse document frequency. The significance of an index term is computed by weighting the occurrence of the term with the importance of the tag associated with it.

Recommending alternate queries during information retrieval is an important feature in a Web-based search engine, because users often do not know the exact terms to locate the information relevant to their interests. Fuzzy ontology is used for query refinement in a domain search engine named Personalizing Abstract Search Service (PASS) [33]. Fuzzy *narrower-* and *broader-than* term relations are defined using a fuzzy conjunction operator. Pruning of redundant relations is made by analyzing the sets of relations involving more than two terms. PASS is used to provide the abstracts of papers from the IEEE Transactions on Neural Networks journal.

8. Text mining

Considering the Web as a huge repository of distributed hypertext, the results from text mining can be included as a topic under Web mining or information retrieval. However we deal with the subject as a separate section, along with image mining in the following section, to highlight their importance in the current context.

A fuzzy decision tree that uses a fuzzy inductive learning to acquire relations from examples is presented in [32]. The authors generate a concept relation dictionary and a classification tree from a random set of daily business reports database of text classes concerning retailing. The algorithm is outlined below.

(i) Allocate a training example set to a new node and stack up the node
(ii) While (pick-up-a-node) do
(iii) Evaluate the allocation of the class to node; if the degree of certainty is acceptable go to step 4
(iv) (a) Create fuzzy sets (fuzzy class items) for each attribute
(b) Calculate evaluation values of the attributes
(c) Select the attribute with the best evaluation value, and allocate the attribute to the node
(d) Decompose the learning set into new subsets, and allocate each subset to a new node
(e) Create new branches, that connect the original and new nodes, and allocate each item to the corresponding branches
(f) Stack up the new nodes and go to step 2

In their experiments the authors use 10,000 evaluation examples, with a decision tree of maximum 90 nodes (37 intermediate and 53 terminal).

A key issue in text mining is keyword extraction. This allows the automatic categorization and classification of text documents. Keyword extraction can be done using clustering methods. relational alternating cluster estimation was used to automatically extract the 20 most relevant keywords from Internet documents in Ref. [31]. Using these keywords, a classification rate of more than 80% could be achieved. The results demonstrate a high efficiency, considering the fact that no additional tools like ontologies were used.

9. Image mining

The literature related to various data mining problems applied to collections of images is large. Fuzzy techniques have been successfully used for information retrieval. A nice overview of activities in this field can be found in [19]. However most authors only refer to a possible application of their algorithm to the fuzzy mining of images from the Web. The heterogeneous structure of the Web is generally composed of multimedia information, viz., text, image (in *jpeg*, *gif* or *mpeg* formats) and sound (in a hypermedia environment). This implies specific variations in clustering or association rule mining, involving different technologies including fuzzy sets. Among the few papers that describe results in direct connection with fuzzy image mining applications from the Web, we mention [7,8].

Video as the format of computer related material is becoming more and more common these days, and many Web pages involve small pieces of movies, video-clips or animation. Fuzzy time related queries are used in [7] to retrieve information inside a video. The queries are handled using Zadeh's principle of *computing with words*, that allows a human-friendly interface. A fuzzy time vocabulary is designed using notions such as time positioning, time descriptors and time relationship. There is a fuzzy partitioning of the universe of discourse for each measure. The authors use the relative temporal relationship framework in order to define the time events. Considering R to be the membership function of a time relationship and T the membership of an event, a new moment/event

is localized in the video by

$$R \circ T(x) = \max_{y} (R(x - y) \wedge T(y)), \tag{11}$$

where $\wedge$ is the fuzzy *And/min* operator. The system is implemented on a Java Search Engine.

Querying for a target image and retrieving it from Web and image databases, based on image similarity, is presented in [8]. A fuzzy c-means algorithm is used to cluster intrinsic image characteristics extracted from subregions of the image. A measure of similarity between pairs of images is determined in terms of the rotation-invariant attributes like color, texture and shape. Color is defined in terms of the hue, saturation, value representation [9] of the average color of the pixel in the region. Texture is represented in terms of the co-occurrence matrices in four directions involving Haralick's parameters [10]. For each database image, the system calculates its attribute matrix and does a whitening transformation before fuzzy clustering to store the representative centroids. When a target image is supplied, the system adopts a similar procedure in order to retrieve the most similar image (in terms of the stored centroid) from the database in response to a query. Experiments are reported on 70 images extracted from the CorelDRAW image database.

10. Conclusions

There is a growing awareness that, in practice, it is easy to discover a huge amount of information from the Web, where most of these patterns are actually obvious, redundant, and useless or uninteresting to the user. To prevent the user from being overwhelmed by a large number of uninteresting patterns, techniques are needed to identify only the useful/interesting patterns and present them to the user.

Fuzzy sets, which constitute the oldest component of soft computing, are suitable for handling the issues related to understandability of patterns, incomplete/noisy data, mixed media information and human interaction, and can provide approximate solutions faster.

In this article we have presented a survey on Web mining involving fuzzy sets and their hybridization with other soft computing tools. The Web mining taxonomy has been described. The individual functions like Web clustering, association rule mining, Web navigation, Web personalization and Semantic Web have been discussed in the fuzzy framework. Finally information retrieval along with text and image mining have been highlighted.

References

[1] J.C. Bezdek, Pattern Recognition with Fuzzy Objective Function Algorithms, Plenum Press, New York, 1981.
[2] D. Camacho, C. Hernandez, J.M. Molina, Information classification using fuzzy knowledge based agents, in: Proc. 10th IEEE Internat. Conf. on Fuzzy Systems, 2001, pp. 4:2575–4:2580.
[3] T.H. Cao, Fuzzy conceptual graph for the semantic web, in: Proc. 2001 BISC Internat. Workshop on Fuzzy Logic and the Internet (FLINT'2001), Berkeley, USA, August 2001.
[4] C.W. Chong, V. Ramachandran, C. Eswaran, Path optimization using fuzzy distance approach, in: Proc. 1999 IEEE International Fuzzy Systems Conf. (FUZZ-IEEE'99), Seoul, Korea, August 1999, pp. III:1771–III:1774.
[5] F. Crestani, G. Pasi, (Eds.), Soft Computing in Information Retrieval: Techniques and Application, vol. 50. Physica-Verlag, Heidelberg, 2000.

[6] R.N. Davé, Characterization and detection of noise in clustering, Pattern Recogn. Lett. 12 (1991) 657–664.
[7] M. Detyniecki, C. Seyrat, R. Yager, Interacting with web video objects, in: Proc. 18th Internat. Conf. of the North American Fuzzy Information Processing Society (NAFIPS'99), 1999, pp. 914–917.
[8] A. Filho, G.L.A. Mota, M.M.B.R. Vellasco, M.A.C. Pacheco, Query by image similarity using a fuzzy logic approach, in: Proc. Fourth Internat. Conf. on Computational Intelligence and Multimedia Applications (ICCIMA 2001), 2001, pp. 389–394.
[9] R.C. Gonzalez, R.E. Woods, Digital Image Processing, Addison-Wesley, Reading, MA, 1992.
[10] R.M. Haralick, K. Shanmugam, I. Dinstein, Textural features for image classification, IEEE Trans. Systems Man Cybernet. 3 (1973) 610–621.
[11] T.-P. Hong, K.-Y. Lin, S.-L. Wang, Mining linguistic browsing patterns in the World Wide Web, Soft Comput. 6 (2002) 329–336.
[12] Y.-P. Huang, Y.-C. Lee, L. Lin, An intelligent approach to mining the related websites, in: Proc. 20th NAFIPS Conf. Vancouver, Canada, July 2001, pp. 435–440.
[13] A. Joshi, R. Krishnapuram, Robust fuzzy clustering methods to support web mining, in: Proc. ACM-SIGMOD Workshop on Data Mining and Knowledge Discovery, August 1998.
[14] A. Joshi, R. Krishnapuram, On mining web access logs, in: Proc. ACM-SIGMOD Workshop on Research Issues in Data Mining and Knowledge Discovery, 2000, pp. 63–69.
[15] K.-J. Kim, S.-B. Cho, A personalized web search engine using fuzzy concept network with link structure, in: Proc. Joint Nineth IFSA World Congress and 20th NAFIPS Internat. Conf. 2001, pp. 1:81–1:86.
[16] R. Krishnapuram, A. Joshi, L. Yi, A fuzzy relative of the k-medoids algorithm with application to document and snippet clustering, in: Proceedings of IEEE International Conference on Fuzzy Systems (FUZZ IEEE'99), Korea, August 1999, pp. 3:1281–3:1286.
[17] R. Krishnapuram, A. Joshi, O. Nasraoui, L. Yi, Low complexity fuzzy relational clustering algorithms for web mining, IEEE Trans. Fuzzy Systems 9 (2001) 595–607.
[18] R. Krishnapuram, J.M. Keller, A possibilistic approach to clustering, IEEE Trans. Fuzzy Systems 1 (1993) 98–110.
[19] R. Kruse, A. Klose, Information mining with fuzzy methods: Trends and current challenges, in: Methods and Models in Automation and Robotics, Szczecin, Poland, 2002, pp. 117–120.
[20] H.-M. Lee, S.-K. Lin, C.-W. Huang, Interactive query expansion based on fuzzy association thesaurus for web information retrieval, in: Proc. the 10th IEEE Internat. Conf. on Fuzzy Systems, 2001, pp. 2:724–2:727.
[21] K.C. Lee, J.S. Kim, N.H. Chung, S.J. Kwon, Fuzzy cognitive map approach to web mining inference amplification, Expert Systems Appl. 22 (2002) 197–211.
[22] M.J. Martin-Bautista, H.L. Larsen, M.A. Vila, A fuzzy genetic algorithm approach to an adaptive information retrieval agent, J. Amer. Soc. Inform. Sci. 50 (1999) 760–771.
[23] G. Meghabghab, Mining user's web searching skills through fuzzy cognitive state map, in: Proc. Joint 9th IFSA World Congress and 20th NAFIPS Internat. Conf. 2001, pp. 1:429–1:434.
[24] S. Mitra, S.K. Pal, P. Mitra, Data mining in soft computing framework: a survey, IEEE Trans. Neural Networks 13 (2002) 3–14.
[25] A. Molinari, G. Pasi, A fuzzy representation of HTML documents for information retrieval systems, in: Proc. Fifth IEEE Internat. Conf. on Fuzzy Systems, 1996, pp. 1:107–1:112.
[26] O. Nasraoui, H. Frigui, R. Krishnapuram, A. Joshi, Extracting web user profiles using relational competitive fuzzy clustering, Internat. J. Artificial Intelligence Tools 9 (2000) 509–526.
[27] A. Nürnberger, A. Klose, Improving clustering and visualization of multimedia data using interactive user feedback, in: Internat. Conf. on Information Processing and Management of Uncertainty in Knowledge Based Systems, Annecy, France, 2002.
[28] W. Pedrycz, Conditional fuzzy c-means, Pattern Recognition Lett. 17 (1996) 625–632.
[29] T.A. Runkler, J. Bezdek, Web mining with relational clustering, Internat. J. Approx. Reason. 32 (2003) 217–236.
[30] T.A. Runkler, J.C. Bezdek, Alternating cluster estimation: a new tool for clustering and function approximation, IEEE Trans. Fuzzy Systems 7 (1999) 377–393.
[31] T.A. Runkler, J.C. Bezdek, Relational clustering for the analysis of internet newsgroups, in: O. Opitz, M. Schwaiger, (Eds.), Exploratory Data Analysis in Empirical Research, Proceedings of the 25th Annual Conference of the German Classification Society, Studies in Classification, Data Analysis, and Knowledge Organization, Springer, Berlin, 2002, pp. 291–299.

[32] S. Sakurai, Y. Ichimura, A. Suyama, R. Orihara, Inductive learning of a knowledge dictionary for a text mining system, in: L. Monostori, J. Vancza, M. Ali (Eds.), Proc. 14th Internat. Conf. Industrial and Engineering Applications of Artificial Intelligence and Expert Systems (IEA/AIE 2001), Lecture Notes in Artificial Intelligence, vol. 2070, Springer, Berlin, 2001, pp. 247–252.

[33] D.H. Widyantoro, J. Yen, Using fuzzy ontology for query refinement in a personalized abstract search engine, in: Proc. Joint Nineth IFSA World Congress and 20th NAFIPS Internat. Conf. Vancouver, Canada, July 2001, pp. 1:610–1:615.

[34] C. Wong, S. Shiu, S.K. Pal, Mining fuzzy association rules for web access case adaptation, in: Proc. Fourth Internat. Conf. on Case-Based Reasoning, Vancouver, Canada, July 2001, pp. 213–220.

[35] L.A. Zadeh, A new direction in AI: towards a computational theory of perceptions, AI Magazine 22 (2001) 73–84.

Journal of Documentation, Vol. 53, No. 2, March 1997

© Aslib, The Association for Information Management.

All rights reserved. Except as otherwise permitted under the Copyright, Designs and Patents Act 1988, no part of this publication may be reproduced, stored in a retrieval system, or transmitted in any form or by any means, electronic, mechanical, photocopying or otherwise without the prior written permission of the publisher.

THESAURUS STRUCTURE DESIGN: A CONCEPTUAL APPROACH FOR IMPROVED INTERACTION

MARÍA J. LÓPEZ-HUERTAS
mjlopez@platon.ugr.es

Facultad de Biblioteconomía y Documentación, Colegio Máximo de Cartuja
Universidad de Granada, 18071 Granada, Spain

The need for thesauri to help users in their search for information in online information systems has been discussed for several decades. Many wide-ranging contributions have been made to solve this problem. Nevertheless, investigation is needed to design a thesaurus structure based on what is relevant for users and generators of information within a specific subject domain. This paper explores the possibility of creating a thesaurus from the cognitive viewpoint. This approach is based on a system (in this case represented by a thesaurus) that organises its representation of knowledge or its classification as closely as possible to the authors' and users' images of the subject domain with the objective of increasing the interaction between users and texts, and thus the communication in a given information retrieval system. From this point of view, the thesaurus structure is considered as the essential foundation on which to base such an interactive thesaurus. Furthermore, this structure is conceived as representing the merging point for both the generators' and the users' models of the subject domain and for their information needs. This paper is dedicated mainly to the generators' side involved in this process. It demonstrates how an author's writings can be used to identify the generators' model and perception of the subject domain, and how these can later be inserted in the thesaurus structure. Discourse analysis is used as a main method to identify the categories and its relevance for building such a structure is discussed. It also outlines a general approach for the user side to set up different methods of getting the users' information needs into the thesaurus structure.

1. INTRODUCTION

THE GROWTH of online information retrieval systems, and users' unassisted searching have made the traditional concept of the indexer-oriented thesaurus obsolete [1], forcing the specialist to find alternatives adapted to this new situation. As a result a new term has been coined, the so-called user-thesaurus, which has attracted much attention. Bates was one of the first in her field to ask for a change in thesaurus construction. She considers the user or end-user thesaurus as opposed to the indexer thesaurus as being designed specifically for online searching [2]. The main features of Bates' user-thesaurus deal with the problem of entry vocabulary, as can be seen from the following list of its most

Aslib, The Association for Information Management
Information House, 20–24 Old Street, London EC1V 9AP
Tel: +44 (0) 171 253 4488, Fax: +44 (0) 171 430 0514
Email: pubs@aslib.co.uk, WWW: http://www.aslib.co.uk/aslib

Journal of Documentation, Vol. 53, No. 2, March 1997

© Aslib, The Association for Information Management.

All rights reserved. Except as otherwise permitted under the Copyright, Designs and Patents Act 1988, no part of this publication may be reproduced, stored in a retrieval system, or transmitted in any form or by any means, electronic, mechanical, photocopying or otherwise without the prior written permission of the publisher.

important characteristics:

1. includes a list of all terms in use in the database;
2. carefully distinguishes terms actually used in a given database from those that are not;
3. provides scope notes for problems likely to be encountered by end users;
4. uses self explanatory names for terms or relationships;
5. includes a vast entry vocabulary, geared to end-user requirements. This feature is particularly important and represents the most noticeable difference between end-user and indexer thesauri [1, p, 369].

At the same time that Bates was developing this area, studies on users' search queries and their behaviour for seeking information were being conducted. This research provided interesting data for thesaurus construction. One of the most important conclusions obtained from these contributions is the way in which thesaurus terminology is conceived. Users' search query terms should be considered as a source for the thesaurus vocabulary in addition to the traditional gathering methods of indexing documents, specialised vocabularies, other indexing languages and experts' contributions. Incorporating the user's terminology as a source for thesauri vocabulary development is a method that is practised now [3, 4], but this was not always the case, or at least it is not explicitly mentioned [5]. In general, the importance of users' terminology is agreed upon by thesaurus theorists and practitioners. The fact that the American standards for thesaurus construction include the so called 'user's warrant' as a source for gathering vocabulary clearly demonstrates this [6].

Another issue of equal importance regarding the newly defined user-thesaurus corresponds to its classification element. Bates mentioned the importance of semantic links in the thesaurus [2, p. 411]. She expressed the need for a network of associations between terms and concepts, much richer and more varied than that found in typical thesauri. Bates described these relationships in the following manner: 'an alphabetical list of terms, legitimate or not, and including word form variations, may be displayed as well as all the conventional thesaurus relationships, narrower, broader and related terms' [1, p. 370].

Strong and Drott [7] pointed out the need for a well structured thesaurus in an article dealing with the design of an end-user thesaurus. They stressed the importance of increasing the thesaurus term relationships in order to create a 'concept space' which is easily browsed upon by users. The authors centred their research on developing a faceted classification for this type of thesaurus, although the criteria used in choosing the facets were not mentioned, nor was there any other sign of linkage between the classification and the user's needs.

Other attempts have been made to build a thesaurus with a more elaborate structure, such as is the thesaurofacet [8] and the clasaurus [9, 10]. The most recent clasaurus, Fugmann's, is thought of as having the qualities and strength of both a thesaurus and a classification, as well as proving efficient for building back-of-the-book indexes: 'the assistance provided by a computerised clasaurus is greatest in the construction of a book index' [10, p. 136]. These thesauri have

Aslib, The Association for Information Management
Information House, 20–24 Old Street, London EC1V 9AP
Tel: +44 (0) 171 253 4488, Fax: +44 (0) 171 430 0514
Email: pubs@aslib.co.uk, WWW: http://www.aslib.co.uk/aslib

Journal of Documentation, Vol. 53, No. 2, March 1997

© Aslib, The Association for Information Management.

All rights reserved. Except as otherwise permitted under the Copyright, Designs and Patents Act 1988, no part of this publication may be reproduced, stored in a retrieval system, or transmitted in any form or by any means, electronic, mechanical, photocopying or otherwise without the prior written permission of the publisher.

a faceted classification to structure the vocabulary, but there is no mention of being based on users' and generators' models.

Few publications on thesaurus structure and classifications have been clearly focused on the user. Pejtersen designed a classification scheme for fiction based on the analysis of user-librarian communication [11]. She also described the construction of a TA thesaurus (Term Association Thesaurus) with a small stock of terms, including 223 controlled key words. The structure of this thesaurus is organised according to users' mental associations with the intention of elaborating an interface for browsing [12].

There is still a need for classification and indexing systems that act in a more user friendly way. As Foskett states: 'there will be a need for a classification and an indexing system which can be used in a more effective and friendly dialogue with users in search for information as in making records of documents at the time of their acquisition in a library' [13].

Many agree that the performance of an IR system can be measured according to the degree of matching between the user's information needs and the texts, depending a great deal upon the coincidence of the models of the participants in the system [14, 15]. Taking this into consideration, it could be concluded that if the model for communication used by the system were close to those of the texts and of the users, the additional effort made by the user in getting to know the system would decrease. In this case, users could achieve their goals using a minimum of cognitive load [16]. Furthermore, it would help in decreasing his or her anxiety [17, pp. 108–127], diminishing the uncertainty of the process. This system would also aid the user in the anomalous state of knowledge he or she is in when seeking information [18]. As a result, the interaction between the users and the system, and the users and the texts would be easier, and therefore it could be increased. Thus this system's model would lead to a more effective information transfer.

In the following sections several points will be discussed: (1) the general framework for building a thesaurus structure; (2) the method used to incorporate the author-text structures into a thesaurus; (3) the criteria established to study the user as a source for a thesaurus structure and an outline for the required research; and (4) the model for designing a thesaurus structure, pointing out the possible outcome.

2. PROPOSED GENERAL FRAMEWORK FOR THESAURUS MODELLING

As the thesaurus structure is the principal point to be considered, vocabulary collecting techniques are not to be discussed here. Nevertheless, it is assumed that users' terminology is included in the thesaurus vocabulary, and that the thesaurus is subject domain oriented. The term structure refers to the conceptual organisations of a particular subject domain, according to the generator-text structures and to the users' information needs. In this case, subject domain encompasses not only a compendium of knowledge but also other features such as tasks, jobs, intentionality etc.

Aslib, The Association for Information Management
Information House, 20–24 Old Street, London EC1V 9AP
Tel: +44 (0) 171 253 4488, Fax: +44 (0) 171 430 0514
Email: pubs@aslib.co.uk, WWW: http://www.aslib.co.uk/aslib

Journal of Documentation, Vol. 53, No. 2, March 1997

© Aslib, The Association for Information Management.

All rights reserved. Except as otherwise permitted under the Copyright, Designs and Patents Act 1988, no part of this publication may be reproduced, stored in a retrieval system, or transmitted in any form or by any means, electronic, mechanical, photocopying or otherwise without the prior written permission of the publisher.

2.1 Cognitive approaches tp thesaurus modelling

The change in the concept of thesaurus design was not an isolated case. It merely represented a minor aspect of a general movement that stressed the need for researchers and professionals in information and library studies to focus their efforts on users' information needs and users' interaction with information systems, as opposed to centring on the system, or on the data alone [19]. This user-oriented philosophy, which has influenced the information and library fields so deeply, is closely related to a new discipline which arose in the mid seventies: cognitive science [20]. Many information science scholars joined this movement, facilitating the transition to a new paradigm in information science, known as the cognitive paradigm [21]. Despite the existence of this general range, there does not exist a unified viewpoint. Belkin outlined the most representative views or foci in this area, and at the same time, he described the common factor 'as being concerned with some sort of human communication system, in which texts play a key role, and of individuals within that system in their interactions with texts (or information), and with one another in relation with such texts' [22].

Harbo *et al.* suggested that 'the task of information retrieval is to bring cognitive structures of authors, system designers and indexers into accord with those of the information worker and the user in order to cope with the actual need' [23]. A similar statement by Ingwersen can be found [24].

Belkin considered that 'the effective transfer of desired information from human generator to human user is the fundamental social problem about which information science is centred' [25]. Later, he mentioned that 'a potentially useful solution to the problem of compatible representation [of need and text] would be to devise ways to represent information need and to then represent information contents of texts in equivalent terms' [25, p. 188]. He also pointed out the advantage of including investigations dealing with the different aspects of this problem under the same framework: the cognitive point of view [25, p. 187].

The objective of this work is to join the authors' structures and the users' structures and needs, denominated by Ingwersen as the user's cognitive space [26], in order to represent the information in the texts as closely as possible to the image that the user might have of his or her information need. In this context, the thesaurus structure is considered the intermediary between both structures – authors' texts and users. This train of thought constitutes a step forward in dispelling a general tendency in indexing language construction, recognised by Stiles when he concluded that 'no subject heading list or thesaurus is rooted in anything other than the say-so of its compilers' [27]. The compilers' main role, in this case, will be to harmonise the structures of both authors' texts and users. This does not mean that the compiler's cognition and knowledge will be excluded, but at least, they will not be the only source.

The thesaurus could be considered as playing a similar role to that of the intermediary of the system. In addition it could be said that its structure or its classification is built according to that of the authors' and users' models. In this case, the thesaurus would encompass the system's above mentioned qualities

Aslib, The Association for Information Management
Information House, 20–24 Old Street, London EC1V 9AP
Tel: +44 (0) 171 253 4488, Fax: +44 (0) 171 430 0514
Email: pubs@aslib.co.uk, WWW: http://www.aslib.co.uk/aslib

Journal of Documentation, Vol. 53, No. 2, March 1997

© Aslib, The Association for Information Management.

All rights reserved. Except as otherwise permitted under the Copyright, Designs and Patents Act 1988, no part of this publication may be reproduced, stored in a retrieval system, or transmitted in any form or by any means, electronic, mechanical, photocopying or otherwise without the prior written permission of the publisher.

increasing the interaction and the communication between human beings and texts as well as users and the knowledge resource in the system. As a consequence, the thesaurus might improve the performance of IR systems. As Harbo *et al.* said 'if an information system operates in a well defined field of knowledge with a stable paradigm, we can expect that authors, indexers, users and reference librarians are likely to have similar images and to work with the same vocabulary, the same concepts and structure. Thus, we can expect a high degree of retrieval success' [23, p. 216]. Although these are hypotheses that have to be tested in real life situations and this article will not come to any definitive conclusions here, there does exist enough theoretical evidence to continue studying the possible configuration of the proposed thesaurus.

2.2 *The thesaurus as an intermediary component in IR*

As Brookes stated, communication between human beings is mediated by a physical channel, and at both ends of the channel cognitive processes take place [28]. Belkin's model of the communication system in information science identifies whose cognition operates at both ends, i.e. that of the generators [of texts] and of the user [25, p. 188]. In other models, the generators are implied as being the authors of the information objects [25, p. 103]. Other cognitive structures are also used in the communication process in IR systems [25, p. 103, 29]. For the purpose of this work, we may consider the third main structure in IR systems as the intermediary – the first and the second ones are those of the generators through their writing (texts) and of the users. If we consider that the thesaurus is the intermediary structure, which we are attempting to design, a schematic model, such as that described by Belkin, seen in Figure 1, is the ideal starting point for the thesaurus structure design.

This model does not represent the connection or the intermediary point between *a* and *b*. This makes the elements that play a major role in IR systems clearer: the user (U), and the generators of information (G) by means of their writings (T). In our opinion, the scheme relates directly to the problem we are trying to address. The design of a thesaurus intended to improve communication in IR systems should consider these elements as a model for building its structure. This thesaurus structure, which takes generators and users into account, simultaneously, could improve interaction between users and texts within the system, given that it will be as close as possible to that of the users and texts. This means that this 'interactive' structure will help the user in getting relevant information from the system knowledge resource.

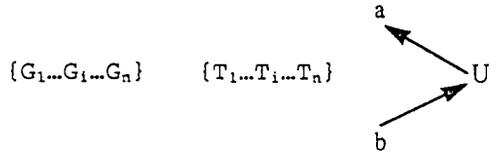


Figure 1. *Model of a communication system in information science* [25, p. 188] *where a represents the user's interrogation of a corpus of texts and b represents the extraction of some texts or some combinations of texts from the corpus*

Aslib, The Association for Information Management
Information House, 20–24 Old Street, London EC1V 9AP
Tel: +44 (0) 171 253 4488, Fax: +44 (0) 171 430 0514
Email: pubs@aslib.co.uk, WWW: http://www.aslib.co.uk/aslib

Journal of Documentation, Vol. 53, No. 2, March 1997

© Aslib, The Association for Information Management.

All rights reserved. Except as otherwise permitted under the Copyright, Designs and Patents Act 1988, no part of this publication may be reproduced, stored in a retrieval system, or transmitted in any form or by any means, electronic, mechanical, photocopying or otherwise without the prior written permission of the publisher.

According to the general viewpoint, it seems obvious that users' information needs must be represented in the thesaurus structure. However, the presence of generators may need further explanation, since they are not explicit in all models. Generators acquire an important role in this design, because they model the texts, and they transfer their knowledge and cognitive structures to their writings. Furthermore, if we know about the generators, we know about the texts or documents.

Some people argue that traditional indexing languages are not very effective because they are not designed for compatible representation of need and text, and they have nothing to say about the generators, which introduce a new element into the system not related to its basic elements [25, p. 189]. The objective of this work is to fashion an indexing language not at all strange to the system's main elements, but rather in harmony with them.

3. GENERATORS' WRITINGS AS A SOURCE FOR THESAURUS STRUCTURE

Generators' writings are one of the main parts in our thesaurus structure design. In this section we intend to show how these writings can be considered as a source for building a thesaurus structure. Writing is a cognitive process [30, 31] that can let us find out not only what the author knows, but also how the topic in the text is treated, how this topic is structured, and what are its relevant components from its author's point of view.

We assumed that there was a previously selected stock of single concept terms, representing a discipline or broad subject area. Each term in this stock will be the unit of analysis and the starting point for building the thesaurus structure. We are interested in knowing what authors write about each of these terms, so we can find out what are the term relevant aspects from their point of view. In theory, any text dealing with the selected terminology would be appropriate. Articles, books, dictionaries etc. about it would be fine.

We started looking up short statements about music terminology. Dictionaries and encyclopedias appeared to be the best sources for that, on the understanding that an author who writes a dictionary of music is a specialist in this field and not a lexicographer. Besides, the information in dictionaries is already condensed, and as it presents some structural patterns it is easier to analyse than longer statements. For example, it is easier to identify relevant aspects regarding one concept, let us say banjo, and criteria for structuring those aspects if we look up definitions for banjo in the dictionary than if we read an article or a book about that concept. Nevertheless, we intend, later, to extend this method to other generators' writings: articles, books etc. The idea is to collect a complete definition for each term, so that we do not miss relevant aspects of it, according to the author's point of view. Likewise, we collected definitions from different authors, so we can see differences, and/or agreement on which are the term relevant aspects for each of these persons. In turn, this can also help in the final evaluation of this stage.

Aslib, The Association for Information Management
Information House, 20–24 Old Street, London EC1V 9AP
Tel: +44 (0) 171 253 4488, Fax: +44 (0) 171 430 0514
Email: pubs@aslib.co.uk, WWW: http://www.aslib.co.uk/aslib

Journal of Documentation, Vol. 53, No. 2, March 1997

© Aslib, The Association for Information Management.

All rights reserved. Except as otherwise permitted under the Copyright, Designs and Patents Act 1988, no part of this publication may be reproduced, stored in a retrieval system, or transmitted in any form or by any means, electronic, mechanical, photocopying or otherwise without the prior written permission of the publisher.

3.1 Searching for relevant criteria for identification of categories: the discourse analysis

We started working with part of the vocabulary stock. A group of terms semantically related was chosen, which was musical instruments. For each term in this group, definitions from different authors, specialists in music, were gathered. Afterwards, each definition was analysed. In the definitions, we were looking for relevant aspects related to the concept being described and for the way in which these aspects were structured, according to the authors' point of view. Part of the authors' intentionality could be identified this way. It was also interesting to look for other semantic information that dictionaries usually give such as the relation between the term sought and other terms in the dictionary, and the genus-species relation that sometimes is included in the definitions. But these two latter elements are not very important at this stage, because the main concern now is not to find out specific relationships among the terms but to identify relevant text elements in the definitions for each term being defined. This would allow us in turn to identify relevant categories for each of them and criteria for organising them into the thesaurus structure, according to the originators' – in this case the music dictionary authors' – point of view.

The general method followed in looking for relevant information in the definitions was to register the meaningful text elements mentioned in the definitions for a particular term. The way of identifying these elements was by using a technique taken from discourse analysis methodology [32, 33]. In particular, the process of creating macrostructures attracted our attention, because macrostructures also deal with text relevance. It is known too that they are often directly expressed in discourse [34]. One of the clues for identifying which text elements are macrorelevant in a writing is their frequency. Van Dijk states that 'by mentioning something repeatedly in the discourse, subjects can be led to believe that this item plays an important role in the macrostructure of the discourse' [34, p. 54]. We selected this technique for the analysis of the definitions. We considered as a text unit all the definitions from different authors for a term. The 'macrostructures' for that term were considered to be those text elements that were mentioned repeatedly in the definitions. It is important, though, to note that the text elements are important as far as they lead us to the identification of relevant categories for the term being defined. So, when we talk about repetition of text elements, we really mean repetition of any expression in the definitions referring to the same category. For instance, if there are text elements related to the morphological/physical aspects of a musical instrument in the definitions from different authors, we say that the morphological/physical category is highly relevant for that instrument and thus, for the thesaurus structure. It does not matter now that one definition talks about strings, shape and size, and others about sound board, neck, body and tuning board. In the end, all these text elements mean morphological/physical characteristics, they are present in every checked definition, and because of that this category is highly relevant. This method also gives the opportunity of ranking the importance of these categories for the term being defined. A statistical account can be done in order to determine such a ranking.

Aslib, The Association for Information Management
Information House, 20–24 Old Street, London EC1V 9AP
Tel: +44 (0) 171 253 4488, Fax: +44 (0) 171 430 0514
Email: pubs@aslib.co.uk, WWW: http://www.aslib.co.uk/aslib

Journal of Documentation, Vol. 53, No. 2, March 1997

© Aslib, The Association for Information Management.

All rights reserved. Except as otherwise permitted under the Copyright, Designs and Patents Act 1988, no part of this publication may be reproduced, stored in a retrieval system, or transmitted in any form or by any means, electronic, mechanical, photocopying or otherwise without the prior written permission of the publisher.

To illustrate what has been said, ten expressions have been chosen: balalaika, banjo, clavichord, harp, lute, mandolin, mandurria, ukelele, violin and violoncello. For each of them, definitions from four dictionaries – A, B, C, D – have been recorded. Relevant text elements in the definitions are italicised.

Balalaika

A: *Russian instrument* with *3 strings tuned in fourths, and plucked with the plectrum, triangular body, like the guitar.* There are *different sizes.* It is played in *folk ensembles* [35].

B: *Stringed plucked instrument*, of *Tartar* origin, *very popular* among the *Russian country people.* It was first mentioned in the reign of *Peter the Great.* It has a *triangular body* with *small holes* in the front, with a *long neck* and *frets, slightly convex sound board.* It has *3 strings.* It was a *soloist instrument* originally, later used in *ensembles (XIX c.)* in six *different sizes* [36].

C: *Russian popular instrument* of the *guitar family.* The *sound board* is *slightly convex*, it has a *long neck* and *3 strings tuned in fourths.* It is played *with the plectrum* [37].

D: A kind of *Russian lute* with *3 strings, finger board with frets* and *triangular body.* It comes in *different sizes.* The *dombra* was the balalaika antecedent, but the former had a longer neck and an oval body [38].

Banjo

A: *Instrument* with *long neck* and *frets*, of *5 to 9 strings plucked with the fingers.* It is somehow *alike the guitar*, but its *body* is *circular*, and with a *stretched membrane* on both *flat sides.* Instrument related to the *black American people.* It was very much used in the *jazz bands* to *keep the rhythm* [35].

B: Favourite *instrument* of *black Americans* who brought it from *Africa*, where it exists under the name of 'bania'. It is a *kind of guitar* with a *long neck* and a *body made of skin* over an *arch, like a timbrel.* It has from *5 to 9 strings.* The first one is *played with the thumb.* The banjo comes in *different sizes.* It plays an important role in the *jazz band*, and because of the popularity of the latter, it has been known in *Europe* [36].

C: A *kind of guitar* with *circular body* on which a *membrane* is *stretched*, acting like a sound box. It has from *5 to 9 strings*, and a *long neck.* Important element in the *New Orleans jazz bands* [37].

D: *Plucked stringed instrument*, a *kind of guitar*, with *6 strings*, although there are also banjos with *5 and 7 strings.* It has a *sweet pitch.* The sound *box* is formed by an *arch* which holds a *stretched membrane fixed by screws* around its circumference. The *back of the body* is also *circular, made of wood*, and it forms a piece with the *neck.* It is played in the *jazz bands* [38].

Clavichord

A: *Rectangular old stringed instrument*, with *strings parallel to the sound board*, with a *38–45 key keyboard.* The strings are *struck* with a *mechanism* that ends in a *key* [35].

B: *Keyboard instrument* originating in the *old plucked medieval instruments.* The *strings* are *struck* by the *tangents.* This is what differentiates the clavichord from the spinet whose strings are plucked with jacks. The *sound board* was *rectangular*, and it was usually placed *on a table* or the like. In *1323 John de Muris* talks about an instrument like the clavichord in his

Aslib, The Association for Information Management
Information House, 20–24 Old Street, London EC1V 9AP
Tel: +44 (0) 171 253 4488, Fax: +44 (0) 171 430 0514
Email: pubs@aslib.co.uk, WWW: http://www.aslib.co.uk/aslib

Journal of Documentation, Vol. 53, No. 2, March 1997

© Aslib, The Association for Information Management.

All rights reserved. Except as otherwise permitted under the Copyright, Designs and Patents Act 1988, no part of this publication may be reproduced, stored in a retrieval system, or transmitted in any form or by any means, electronic, mechanical, photocopying or otherwise without the prior written permission of the publisher.

work '*Musica especulativa*', and *William Horwood* was teaching this instrument in *1477*. In the *XV century* is *well known in Europe*. It is often cited in the court chronicles of *Aragón y Castille*. *A. Cabezón* wrote music for this instrument, and he also played it. The clavichord was almost exclusively used for *domestic music* and the *accompaniment of voices*. Compositions for this instrument are few. An important piece was the '*Neue Clavierübung*' by *Johan Kuhnau* [36].

C: *Stringed instrument struck* by a *mechanism* actioned by a *keyboard* [37].

D: *Instrument* that has *already disappeared in today's music*. It had a great number of *strings* that were *struck by a keyboard mechanism*. The improved version of this instrument gave place to the modern *piano* [38].

Harp

A: *Triangular instrument* with *47 uneven length strings, plucked with the fingers*. Today it embraces *six octaves and a half*. The standard harp has its strings *diatonically tuned in C flat major*. To play any *chromaticism* the *pedals* are hit, and they make any note a semitone or a tone higher. The harp has *7 pedals* [35].

B: *Plucked stringed instrument*, consisting of *one or more rows of strings* of *different length*. It is one of the *oldest instruments known* until the *Christian era*, and its shape was much like the modern instrument. It was used in the old *Egyptian empire, Persia, India, Turkey*. In *Greece* and *Rome* was not used very much. The present harp is *tuned in C major*. This is a *diatonic instrument*. *Glissando effects* can be obtained from it. The main harp composers are *Mozart, Marin, Krumpholtz, Dalvimare, Godefroid, Bochsa, Debussy, Ravel*, etc. [36].

C: *Triangular shaped instrument*. It has *uneven length strings (fixed* at the end to a *sound board)* that are *plucked with the fingers* of *both hands*. It is a *diatonic instrument tuned in C flat*. [37].

D: *Stringed plucked instrument without neck*. It is *diatonic tuned in C flat*. It has *7 pedals* and *47 strings*. Each pedal is for a note in the musical scale [38].

Lute

A: *Stringed plucked instrument*. It comes from *Oriental countries*, brought to the *Occident* by the *Arabs* and the *Crusaders*. *Commonly used* in the *Middle Ages*. It was the *favourite instrument* for *domestic music* in the *XVI–XVII centuries*. For this reason it has a large repertoire. It has *half a pear shape*, the *neck* with *frets*, a *rose in the front*, the *tuning board bent backward*, with a *different number of strings* (*11* as a rule, *tuned re, la, fa, doh, sol*). It has a *finger touch notation*. Most of the *sacred and secular polyphonic music was arranged* for the lute [35].

B: *European stringed instrument with neck* and *tuning board, convex body* and *sound board* with a *central rose*. Straight *tuning board bent backward*. The *neck* is *wide* with *9–10 frets*. In the *XV century* it was played either *with the fingers* or *with the plectrum* but afterwards it has been a *plucked instrument*. It comes from the *Arab-Persian ûd* that came to *Europe* when it was invaded by the *Arabs* through *Spain* and *Sicily*. The lute has had a place in *Occidental music* since the *XIV century*. In the *XV century*, it played an *important role* in musical life. It was very much played in *ensembles*. It was used for *replacing absent voices* and for *accompaniment*. The lute has *11 strings* (5 double and 1 single). The lute *schools of music* made *variations* and *arrangements* of *sacred and secular polyphonic music* for this instrument. The main schools were located in *Spain, Italy, France* and

Aslib, The Association for Information Management
Information House, 20–24 Old Street, London EC1V 9AP
Tel: +44 (0) 171 253 4488, Fax: +44 (0) 171 430 0514
Email: pubs@aslib.co.uk, WWW: http://www.aslib.co.uk/aslib

Journal of Documentation, Vol. 53, No. 2, March 1997

© Aslib, The Association for Information Management.

All rights reserved. Except as otherwise permitted under the Copyright, Designs and Patents Act 1988, no part of this publication may be reproduced, stored in a retrieval system, or transmitted in any form or by any means, electronic, mechanical, photocopying or otherwise without the prior written permission of the publisher.

Germany [36].

C: *Stringed plucked instrument* with a *sound board* and a *neck* with *moving frets*. The *sound board is convex* and *potbellied in the back*. The *tuning board* is *bent backward* making a right angle. It had a *very important role* in the *instrumental music* of the *XVI–XVII centuries* because of its *polyphonic possibilities* [37].

D: *Old stringed instrument plucked with the plectrum*. It has *6 double strings* [38].

Mandolin

A: The *most recent instrument* in the *lute family*. It is very much used in the *South of Italy*. It appeared in *Europe* between the *XVI* and *XVII centuries* when the *mandola* disappeared. It is a *plucked stringed instrument*. It is played with the *plectrum*, and it has a *convex pear shape*. It is *smaller than the lute*. There are two kinds of mandolin, the *Neapolitan mandolin* and the *Milanese mandolin* [35].

B: *Instrument* of *uncertain origins*, that appeared in *Europe* in the second half of the *XVIII century*, specially made by *Neapolitan makers*. The mandolin has kept *popularity* through time in *popular music*. It is in *common use* in *Italy* specially in *Naples* as a *melodic instrument*. It is a *stringed instrument plucked with the plectrum*, with a *convex sound board, pear shaped, more convex than the lute, but smaller*. There were two kinds of mandolins, the *Neapolitan mandolin* and the *Milanese mandolin*. *Gretry* composed the serenade *'L'amant jaloux'* for mandolin. *Pasiello's 'The Barber of Seville'* and *Mozart's 'Don Giovanni'* are pieces for the mandolin. See the *methods by Cottin, Pietrapertosa, Bartholuzzi, Köler*, etc. [36].

C: *Instrument similar to the lute but smaller*. It has *4 double strings tuned like those of the violin, plucked with the plectrum*, and it has a *convex sound board* [37].

D: *Stringed instrument plucked with the plectrum, 4 double strings tuned like those of the violin* [38].

Mandurria (Mandola)

A: A *common instrument in the Iberian Peninsula*. It looks *like the lute, smaller*, with *6 double strings plucked with the plectrum, tuned in fourths*. It is very much used in *student music groups* and *serenade groups* [35].

B: *Stringed popular instrument smaller than the guitar*. It consists of a *circular sound board* and a *short neck* with *14 metal frets*, the *back of the body is flat*. It has *6 double strings*, and it is played with the *plectrum*. It is a *very popular instrument in Spain* where there are *ensembles* that use this instrument [36].

C: *Typical Spanish instrument*. Its shape is *similar to that of the lute, but much smaller* and with a *flat body*. It has *6 double strings tuned in fourths*. It is part of *student music groups* and *serenade groups* [37].

D: *Small instrument* with *6 double strings* and *narrow sound board*. It is a *typical Spanish instrument, like the lute*, used for playing *popular music* [38].

Ukelele

A: *Guitar* from *Hawaii*. A *small instrument* with *4 strings* and *long finger board, with or without frets*, with a *finger touch notation*. It was *popular* in the *USA* during the *1920s*, mainly in *popular music ensembles* [35].

B: *Small modern guitar* used in the *Hawaiian islands*, with *4 strings*. It comes from a similar instrument called *machete* in Portuguese. It was a *very*

Aslib, The Association for Information Management
Information House, 20–24 Old Street, London EC1V 9AP
Tel: +44 (0) 171 253 4488, Fax: +44 (0) 171 430 0514
Email: pubs@aslib.co.uk, WWW: http://www.aslib.co.uk/aslib

Journal of Documentation, Vol. 53, No. 2, March 1997

© Aslib, The Association for Information Management.

All rights reserved. Except as otherwise permitted under the Copyright, Designs and Patents Act 1988, no part of this publication may be reproduced, stored in a retrieval system, or transmitted in any form or by any means, electronic, mechanical, photocopying or otherwise without the prior written permission of the publisher.

popular instrument in the *USA* in the *1920s*, and it was included in *jazz ensembles*. The music for this instrument uses a *notation* similar to that used by the '*lute players*' in the *XVI century* [36].

C: *Small guitar* with *4 strings* from *Indonesia*. It uses a *finger touch notation* [37].

D: *Small guitar* of *4 strings* that comes from the *Portuguese machete*. It appeared in *Hawaii* in the *XIX century*. It is used for *dance music*, and it is very much related to the *banjo* [38].

Violin

A: Today it is *the most well known bowed stringed instrument in the world*. The violin and the *instruments belonging to the violin family have eliminated all other bowed stringed instruments*. We do not think that it is necessary here either to describe the instrument or enumerate its parts. *Its tuning is sol, re, la, me*. The violin appeared at the end of the *XVI century* as the *soprano instrument* of the *viol family* bearing by this time the names of *violin piccolo* or *violino* [35].

B: *Main bowed stringed instrument of the bowed stringed family*. It has *4 strings tuned sol, re, la, me. It sides are slightly convex*. It has a *neck without frets* that ends in a scroll with a *bridge*. The *strings are scraped* with a *bow*, and rarely it is plucked with the fingers (*pizzicato*). It is a *melodic instrument*. In the *orchestra* there are two groups: *first violins* and *second violins*. It comes from the *soprano instrument* of the *viol family (XVI c.)* through the *violino (1560)* and the *violin piccolo (XVII c.)*. Its *virtuosity* started in the *XVI century* for reproducing *imitative effects* and *dancing techniques*. In the second half of the *XVI century*, the violin acquired its definite look thanks to *Legrenzi, Vitali, Torelli and Corelli*. It came to the forefront with *Vivaldi, Veracini, Tartini, Locatelli, Leclair and Germiniani*, and later with *Viotti, Kreutzer, Graun, Stamitz, Spohr, Paganini, Joachim and Sarasate*. The violin was perfected by *Gasparo da Salò, Amati, Stradivarius, Guarnieri*. The violin is *essential* in *orchestras* and *chamber ensembles*. It is also one of the *main soloist instruments* [36].

C: *Bowed stringed instrument*, the *soprano* of the *strings section* in the modern *orchestra*. It has a *sound board* formed by *two slightly convex sides*. It has *purfling* on the sides and *f shaped sound holes on the belly*. Between them, the *bridge* is located. The *neck* comes from the *sound board*. The *strings (4 tuned sol, re, la, me)* are *fixed* at the *sound board* in its lower part and at the *tuning pins* at the top of the neck in its upper part [37].

D: *Bowed stringed instrument*. It has *4 strings tuned sol re la me*, and it is *the smallest* and *the highest pitched* in the *violin family*. The first violins were made in *Italy* in the *XVI century*. Its predecessors were the *rebec, the lira da braccio* and *the viella*. The most famous *violin makers* were *Amati, Guarnieri and Stradivarius*. The violin has a long tradition in *folk music* [38].

Violoncello

A: *Bowed stringed instrument, twice the violin size*, with a *704 mm string length*. It is played *between the knees on the floor*. It has *4 strings tuned doh, sol, re, la*. Its music is written in three registers *bass clef, treble clef* and *c clef*. It has a *deep sonority*. It appeared at the middle of the *XVI century* as the *bass* of the *viol da braccio family*, under the name of *viola*. Its actual name comes from the *XVII century*. It was the *bass continuum accompaniment* as the solo was played by the *viol da gamba*. In *1700*, the violoncello was

Aslib, The Association for Information Management
Information House, 20–24 Old Street, London EC1V 9AP
Tel: +44 (0) 171 253 4488, Fax: +44 (0) 171 430 0514
Email: pubs@aslib.co.uk, WWW: http://www.aslib.co.uk/aslib

Journal of Documentation, Vol. 53, No. 2, March 1997

© Aslib, The Association for Information Management.

All rights reserved. Except as otherwise permitted under the Copyright, Designs and Patents Act 1988, no part of this publication may be reproduced, stored in a retrieval system, or transmitted in any form or by any means, electronic, mechanical, photocopying or otherwise without the prior written permission of the publisher.

improved by *Stradivarius*, and started to be a *solo instrument (Jacquini, Vivaldi, Dall' Abaco, L. Leo, Tartini, G. Mathias Monn, Bach, Haydn, Bocherini, Corrette's method)*. It took the viol da gamba's place in *1800*. Many scores were written for the violoncello in the *XIX* and *XX centuries*. *Concertos by Schumann, Dvorak, Volkmann, Prokofiev, Shostakovich, Hindemith; sonatas by Beethoven, Brahms, Fauré, R. Strauss, Rachmaninoff, Debussy, Britten, Prokofiev*, etc. The performers who improved playing techniques on violoncello were *Franciscello, Dall' Abaco, Lauretti, Duport, Romberg, Casals, Maréchal, Cassadó*, etc. [35].

B: This *instrument* appeared in the middle of the *XVI century* as the *bass of the viol da braccio family*, under the name of *viola*. By the *XVII*, it was named *violoncino* or *violoncello*. Its role was to *accompany the bass continuum*, because the *viol da gamba* was preferred to it. From 1700 to 1725 the violoncello was improved by *Stradivarius*, taking its actual shape. It has *4 strings tuned doh sol re la*. Its music is written in *bass clef, treble clef and c clef*, and it has a range of *three octaves and one fifth*. It is mainly a *singing instrument* whose registers recall those of the human voice. It was introduced in the *orchestra* by the Italian *Bastitini. Corelli* and *Tartini* took with them violoncello players who played the *bass* in the *sonatas*. Among the first violoncello players were *Franciscello, Bertaut, Janson, de Cupis, de Duport*, etc. The violoncello *ousted the viol da gamba*, and in the *XVIII century* started its success. Most of the music composers have written scores for this instrument, giving to it a *main role* for *instrumental and dramatic music. Beethoven, Haydn, Handel, Bach, Mozart* wrote compositions for violoncello. *Haydn and Mozart* used it very much for their *chamber music*. The music for *solo violoncello* started with the *'Ricercari'* by *Dom Gabrielli* and the *sonatas* by *Jacchini, Lanzetti, Cirri, Cervetto*, etc. The violoncello is held *between the legs* like the *viol da gamba*, and both use the *pizzicatto technique*. *Duport* has developed the violoncello playing technique. Among noted violoncello players are *Abaco, Berteau, Antoniotti, Bocherini, Cervetto, Duport, Casals*, etc. The most famous *learning methods* are those by *Mich, Corrette, de Cupis, Duport*, etc. [36].

C: *Bowed instrument* belonging to the *violin family*, but it has a *low register* because of its *bigger size* [37].

D: *Bowed stringed instrument* with *4 strings* [38].

Analyses of the above definitions for the ten musical instruments identify categories for each of the text elements italicised in the definitions. Figure 2 lists both features – text elements and categories and they are enumerated under the instrument they belong to. Asterisks mean that (i) the text elements are inferred from the definitions; (ii) categories will be named when more data are at hand.

Almost all the relevant text elements in Figure 2 keep the text wording, because they are expressed in the definitions, but, sometimes, non-explicit relevant elements need to be inferred from the text, as is the case of popular instruments in the sample group, which is inferred from the expression 'favourite instrument of black Americans' and jazz music from 'jazz band' in Banjo; popular music from 'popular instrument' and folk music from 'folk ensembles' in Balalaika; vocal music from 'accompaniment of voices' in Clavichord and scraped from 'bowed stringed instrument' in Violoncello. It can be seen, too, that new relevant text elements merge from checking different authors. This is the case of expressions such as 'to keep the rhythm', 'made of

Aslib, The Association for Information Management
Information House, 20–24 Old Street, London EC1V 9AP
Tel: +44 (0) 171 253 4488, Fax: +44 (0) 171 430 0514
Email: pubs@aslib.co.uk, WWW: http://www.aslib.co.uk/aslib

Journal of Documentation, Vol. 53, No. 2, March 1997

© Aslib, The Association for Information Management.

All rights reserved. Except as otherwise permitted under the Copyright, Designs and Patents Act 1988, no part of this publication may be reproduced, stored in a retrieval system, or transmitted in any form or by any means, electronic, mechanical, photocopying or otherwise without the prior written permission of the publisher.

Balalaika		like the guitar	relation with other instruments/*
		like the lute*	relation with other instruments/*
Relevant text elements	*Categories*	guitar family	relation with other instruments/family
		dombra	relation with other instruments/origins
instrument	type of sound device/solo		
stringed	source of sound	Banjo	
plucked	manner or producing sound		
plectrum	means of producing sound	*Relevant text elements*	*Categories*
3 strings	physical/morphological characteristics		
triangular [body]	physical/morphological characteristics	instrument	type of sound device/solo
body	physical/morphological characteristics	stringed	source of sound
finger board	physical/morphological characteristics	plucked	manner of producing sound
neck	physical/morphological characteristics	with the fingers	means of producing sound
long [neck]	physical/morphological characteristics	long [neck]	physical/morphological characteristics
with frets	physical/morphological characteristics	frets	physical/morphological characteristics
convex [sound board]	physical/morphological characteristics	circular [body]	physical/morphological characteristics
holes [front of the body]	physical/morphological characteristics	membrane	physical/morphological characteristics
different sizes	physical/morphological characteristics	flat sides	physical/morphological characteristics
soloist instrument	musical functions/*	arch	physical/morphological characteristics
accompaniment instrument	musical functions/*	5-9 strings	physical/morphological characteristics
tuned in fourths	musical character	different sizes	physical/morphological characteristics
popular instrument	social functions	fixed by screws	physical/morphological characteristics
ensembles	musical functions/ensembles	box	physical/morphological characteristics
folk ensembles	musical functions/ensembles	stretched [membrane]	physical/morphological characteristics
folk music*	musical functions/genre	neck	physical/morphological characteristics
popular music*	musical functions/genre	strings	physical/morphological characteristics
Tartars	ethnic characteristics	made of skin	material
country people	ethnic characteristics	made of wood	material
Russia	place	sweet pitch	musical character
Peter the Great	time	to keep the rhythm	musical functions/*
XIX century	time	jazz bands	musical functions/ensembles

Figure 2. *Analysis of text elements*

Aslib, The Association for Information Management
Information House, 20–24 Old Street, London EC1V 9AP
Tel: +44 (0) 171 253 4488, Fax: +44 (0) 171 430 0514
Email: pubs@aslib.co.uk, WWW: http://www.aslib.co.uk/aslib

Journal of Documentation, Vol. 53, No. 2, March 1997

© Aslib, The Association for Information Management.

All rights reserved. Except as otherwise permitted under the Copyright, Designs and Patents Act 1988, no part of this publication may be reproduced, stored in a retrieval system, or transmitted in any form or by any means, electronic, mechanical, photocopying or otherwise without the prior written permission of the publisher.

played with the thumb	playing techniques	court music	musical functions/genre
jazz music	musical functions/genre	vocal music*	musical functions/genre
popular instrument	social functions	old instrument	time
black people	ethnic characteristics	Middle Ages	time
Africa	place	XV century	time
USA	place	XIV century	time
New Orleans	place	Europe	place
Europe	place	Castille	place
like the guitar	relation with other instruments*	Aragon	place
like the timbrel	relation with other instruments*	old plucked string instruments	relation with other instruments/origins
		spinet	relation with other instruments/*
Clavichord		piano	relation with other instruments/influences
Relevant text elements	*Categories*		musical functions/status
		disappeared today	composers
instrument	type of sound device/solo	Cabezón	composers
stringed*	source of sound	Johan Kuhnau	theoreticians
struck	manner of producing sound	Juan de Muris	compositions
mechanism	means of producing sound	'Neue Clavierübung'	works on music
keyboard	means of producing sound	Musica speculativa	
rectangular shape	physical/morphological characteristics		
strings	physical/morphological characteristics	Harp	
sound board	physical/morphological characteristics		
strings parallel to the sound board	physical/morphological characteristics	*Relevant text elements*	*Categories*
38-45 keys	physical/morphological characteristics	instrument	type of sound device/solo
tangents	physical/morphological characteristics	stringed	source of sound
on a table	playing techniques	plucked	manner of producing sound
accompaniment		with the fingers	means of producing sound
instrument	musical functions/*	triangular shape	morphological/physical characteristics
domestic music	musical functions/genre	sound box	morphological/physical characteristics

Figure 2. *Analysis of text elements*

Aslib, The Association for Information Management
Information House, 20-24 Old Street, London EC1V 9AP
Tel: +44 (0) 171 253 4488, Fax: +44 (0) 171 430 0514
Email: pubs@aslib.co.uk, WWW: http://www.aslib.co.uk/aslib

Journal of Documentation, Vol. 53, No. 2, March 1997

© Aslib, The Association for Information Management.

All rights reserved. Except as otherwise permitted under the Copyright, Designs and Patents Act 1988, no part of this publication may be reproduced, stored in a retrieval system, or transmitted in any form or by any means, electronic, mechanical, photocopying or otherwise without the prior written permission of the publisher.

strings fixed	morphological/physical characteristics	stringed	source of sound
7 pedals	morphological/physical characteristics	plucked	manner of producing sound
47 strings	morphological/physical characteristics	with the fingers	means of producing sound
without neck	morphological/physical characteristics	with the plectrum	means of producing sound
1 or more rows [strings]	morphological/physical characteristics	neck	morphological/physical characteristics
uneven length [strings]	morphological/physical characteristics	tuning board	morphological/physical characteristics
with both hands	playing techniques	backward bent [tuning board]	morphological/physical characteristics
glissando effects	playing techniques	body	morphological/physical characteristics
tuned C♮ major	musical character	6 double strings	morphological/physical characteristics
diatonic instrument	musical character	11 strings	morphological/physical characteristics
6 ½ octave range	musical character	half a pear shape	morphological/physical characteristics
chromaticism	musical character	with frets	morphological/physical characteristics
Egypt	place	9-10 frets	morphological/physical characteristics
Persia	place	moving [frets]	morphological/physical characteristics
India	place	2 central roses [front]	morphological/physical characteristics
Turkey	place	convex [back of the body]	morphological/physical characteristics
Greece	place	potbellied [back of the body]	morphological/physical characteristics
Rome	place	finger touch notation	musical character
Christian Era	time	tuned re, la, fa, doh, sol	musical character
Mozart	composers	polyphonic possibilities	musical character
Marin	composers	variations	musical function/arrangements
Krumpholtz	composers	arrangements	musical function/arrangements
Dalvimare	composers	replaces absent voices	musical functions/*
Godefroid	composers	accompaniment instrument	musical functions/*
Debussy	composers	domestic music	musical functions/genre
Ravel	composers	instrumental music	musical functions/genre
		secular polyphonic music	musical functions/genre
Lute		sacred polyphonic music	musical functions genre
Relevant text elements	*Categories*	Occidental music	musical functions/genre
		ensembles	musical functions/ensembles
instrument	type of sound device/solo	favourite instrument	social functions

Figure 2. *Analysis of text elements*

Aslib, The Association for Information Management
Information House, 20–24 Old Street, London EC1V 9AP
Tel: +44 (0) 171 253 4488, Fax: +44 (0) 171 430 0514
Email: pubs@aslib.co.uk, WWW: http://www.aslib.co.uk/aslib

Journal of Documentation, Vol. 53, No. 2, March 1997

© Aslib, The Association for Information Management.

All rights reserved. Except as otherwise permitted under the Copyright, Designs and Patents Act 1988, no part of this publication may be reproduced, stored in a retrieval system, or transmitted in any form or by any means, electronic, mechanical, photocopying or otherwise without the prior written permission of the publisher.

schools of music	teaching centres	neck*	morphological/physical characteristics
Spain	place	pear shaped	morphological/physical characteristics
France	place	smaller than the lute	morphological/physical characteristics
Italy	place	melodic instrument	musical characteristics
Germany	place	tuned sol re la me	musical characteristics
Persia	place	popular instrument	social functions
Europe	place	popular music	musical functions/genre
Occident	place	Europe	place
Oriental countries	place	Naples	place
Arab people	ethnic characteristics	Italy	place
Crusaders	religious/military groups	makers	instrument's makers
XIV-XVII centuries	time	XVII-XVIII centuries	time
Middle Ages	time	lute family	relation with other instruments/family
Ancient times	time	like the lute	relation with other instruments/*
important role	significance-role	Neapolitan mandolin	relation with other instruments/*
ūd	relation with other instruments/origins	Milanese mandolin	relation with other instruments/*
		unclear origins	relation with other instruments/origins
Mandolin		Gretry	composers
		Pasiello	composers
	Categories	Mozart	composers
Relevant text elements		Schoemberg	composers
instrument	type of sound device/solo	Mahler	composers
stringed	source of sound	Handel	composers
plucked	manner of producing sound	'L'amant jaloux'	compositions
with plectrum	means of producing sound	'The Barber of Seville'	compositions
4 strings	morphological/physical characteristics	'Don Giovanni'	compositions
double stringed	morphological/physical characteristics	Cottin	learning methods
convex	morphological/physical characteristics	Pietrapertosa	learning methods
[back of the body]		Bartoluzzi	learning methods
convex shape	morphological/physical characteristics	Köhler	learning methods

Figure 2. *Analysis of text elements*

Aslib, The Association for Information Management
Information House, 20–24 Old Street, London EC1V 9AP
Tel: +44 (0) 171 253 4488, Fax: +44 (0) 171 430 0514
Email: pubs@aslib.co.uk, WWW: http://www.aslib.co.uk/aslib

Journal of Documentation, Vol. 53, No. 2, March 1997

© Aslib, The Association for Information Management.

All rights reserved. Except as otherwise permitted under the Copyright, Designs and Patents Act 1988, no part of this publication may be reproduced, stored in a retrieval system, or transmitted in any form or by any means, electronic, mechanical, photocopying or otherwise without the prior written permission of the publisher.

Mandurria (Mandola)		metal [frets]	material
		Ukelele	
Relevant text elements	*Categories*	*Relevant text elements*	*Categories*
instrument	type of sound device/solo	instrument*	type of sound device/solo
stringed	source of sound	stringed*	source of sound
plucked	manner of producing sound	plucked*	means of producing sound
with plectrum	means of producing sound	4 strings	morphological/physical characteristics
6 double strings	morphological/physical characteristics	with frets	morphological/physical characteristics
small instrument	morphological/physical characteristics	without frets	morphological/physical characteristics
circular [sound board]	morphological/physical characteristics	long [finger board]	morphological/physical characteristics
sound board	morphological/physical characteristics	finger board	morphological/physical characteristics
neck	morphological/physical characteristics	neck*	morphological/physical characteristics
very short [neck]	morphological/physical characteristics	small [instrument]	morphological/physical characteristics
14 frets	morphological/physical characteristics	finger touch notation	notation
flat body	morphological/physical characteristics	popular ensembles	musical functions/ensembles
flat [back of the body]	morphological/physical characteristics	jazz ensembles	musical functions/ensembles
tuned in fourths	musical character	jazz music*	musical functions/genre
ensembles	musical functions/ensembles	popular music	musical functions/genre
serenade groups	musical functions/ensembles	dance music	musical functions/genre
student music groups	musical functions/ensembles	popular instrument	social functions
serenades	musical functions/genre	Portugal	place
passacaglia*	musical functions/genre	USA	place
instrumental music*	musical functions/genre	Indonesia	place
popular music	musical functions/genre	Hawaii	place
popular/typical instrument	social functions	1920s	time
Spain	place	XIX century	time
Iberian Peninsula	place	modern	time
Aragon	place	guitar	relation with other instruments/*
like the lute	relation with other instruments/*	machete	relation with other instruments/*
like the guitar	relation with other instruments/*		
made out of gut [strings]	material		

Figure 2. *Analysis of text elements*

Aslib, The Association for Information Management
Information House, 20–24 Old Street, London EC1V 9AP
Tel: +44 (0) 171 253 4488, Fax: +44 (0) 171 430 0514
Email: pubs@aslib.co.uk, WWW: http://www.aslib.co.uk/aslib

Journal of Documentation, Vol. 53, No. 2, March 1997

© Aslib, The Association for Information Management.

All rights reserved. Except as otherwise permitted under the Copyright, Designs and Patents Act 1988, no part of this publication may be reproduced, stored in a retrieval system, or transmitted in any form or by any means, electronic, mechanical, photocopying or otherwise without the prior written permission of the publisher.

banjo	relation with other instruments/*	orchestra	musical functions/orchestras
		chamber ensembles*	musical functions/ensembles
Violin		chamber music	musical functions/genre
		dance music	musical functions/genre
	Categories	Italy	place
		XVI century	time
		XVII century	time
instrument	type of sound device/solo	violin family	relation with other instruments/family
stringed	source of sound	rebec	relation with other instruments/origins
scraped	manner of producing sound	lira de braccio	relation with other instruments/origins
bowed	means of producing sound	vielle	relation with other instruments/origins
4 strings	morphological/physical characteristics	viol family	relation with other instruments/origins
neck	morphological/physical characteristics	known in the whole word	musical functions/significance-role
sound board	morphological/physical characteristics	main instrument	musical functions/significance-role
ears	morphological/physical characteristics	essential in	musical functions/significance-role
f shaped [sound holes]	morphological/physical characteristics	orchestras/ensembles	
without frets	morphological/physical characteristics	main soloist instrument	musical functions/significance-role
belly [sound board]	morphological/physical characteristics	eliminates all other	musical functions/significance-role
tuning pins	morphological/physical characteristics	bowed instruments	
purfling	morphological/physical characteristics	Vivaldi	composers
slightly convex sides	morphological/physical characteristics	Veracini	composers
bridge	morphological/physical characteristics	Tartini	composers
pizzicato	playing techniques	Locatelli	composers
virtuosity	playing techniques	Leclair	composers
tuned sol, re, la, me	musical character	Germiniani	composers
highest pitches	musical character	Viotti	composers
melodic instrument	musical character	Kreutzer	composers
soprano instrument	musical character	Graun	composers
solo instrument	musical functions/*	Stamitz	composers
imitative effects	musical functions/*	Spohr	composers
dance techniques	musical functions/*	Paganini	composers
first violin	musical functions/orchestras	Joachim	composers
second violin	musical functions/orchestras		

Figure 2. *Analysis of text elements*

Aslib, The Association for Information Management
Information House, 20-24 Old Street, London EC1V 9AP
Tel: +44 (0) 171 253 4488, Fax: +44 (0) 171 430 0514
Email: pubs@aslib.co.uk, WWW: http://www.aslib.co.uk/aslib

Journal of Documentation, Vol. 53, No. 2, March 1997

© Aslib, The Association for Information Management.

All rights reserved. Except as otherwise permitted under the Copyright, Designs and Patents Act 1988, no part of this publication may be reproduced, stored in a retrieval system, or transmitted in any form or by any means, electronic, mechanical, photocopying or otherwise without the prior written permission of the publisher.

Sarasate	composers	chamber orchestra	musical functions/orchestras
Gasparo da Salò	instrument makers	instrumental music	musical functions/genre
Amati	instrument makers	dramatic music	musical functions/genre
Stradivarius	instrument makers	chamber music	musical functions/genre
Guarnieri	instrument makers	sonatas	musical functions/forms
		concertos	musical functions/forms
Violoncello		ricercares	musical functions/forms
		held between legs/knees	playing techniques
Relevant text elements	*Categories*	on the floor	playing techniques
		pizzicato	playing techniques
instrument	type of sound device/solo	viol da braccio family	relation with other instruments/origins
stringed	source of sound	bass continuum	relation with other instruments/*
scraped	manner of producing sound	viol da gamba	relation with other instruments/*
bowed	means of producing sound	violin family	relation with instruments/family
4 strings	physical/morphological characteristics	main role in instrumental music	musical functions/significance-role
704 mm. string length	physical/morphological characteristics	XVI century	time
twice the violin size	physical/morphological characteristics	XVII century	time
bigger size	physical/morphological characteristics	XVIII century	time
singing instrument	musical character	Batistini	performers
tuned doh sol re la	musical character	Corelli	performers
low registers	musical character	Tartini	performers
deep sonority	musical character	Franciscello	performers
three octaves and one fifth range	musical character	Bertaut	performers
bass clef on 4th line	musical functions/scores	Janson	performers
treble clef	musical functions/scores	Duport	performers
c clef	musical functions/scores	Tilliere	performers
solo instrument	musical functions/*	Antoniotti	performers
bass [viol fam.]	musical functions/*	Bocherini	performers
reinforce the bass continuum	musical functions/*	Servais	performers
accompaniment instrument	musical functions/*	Cervetto	performers
orchestra	musical functions/orchestras	Casals	performers

Figure 2. *Analysis of text elements*

Aslib, The Association for Information Management
Information House, 20–24 Old Street, London EC1V 9AP
Tel: +44 (0) 171 253 4488, Fax: +44 (0) 171 430 0514
Email: pubs@aslib.co.uk, WWW: http://www.aslib.co.uk/aslib

Journal of Documentation, Vol. 53, No. 2, March 1997

© Aslib, The Association for Information Management.

All rights reserved. Except as otherwise permitted under the Copyright, Designs and Patents Act 1988, no part of this publication may be reproduced, stored in a retrieval system, or transmitted in any form or by any means, electronic, mechanical, photocopying or otherwise without the prior written permission of the publisher.

Maréchal	performers	Shostakovich	composers
Cassadó	performers	Saëns	composers
Beethoven	composers	Fauré	composers
Handel	composers	Ibert	composers
Bach	composers	R. Strauss	composers
Haydn	composers	Reger	composers
Mozart	composers	Debussy	composers
Dom Gabrieli	composers	Casella	composers
Jacchini	composers	Britten	composers
Lanzetti	composers	Rachmaninoff	composers
Cirri	composers	Stradivarius	instrument maker
Cervetto	composers	Mich	learning methods
Vivaldi	composers	Corrette	learning methods
Schumann	composers	de Cupis	learning methods
Dvorak	composers	Duport(method)	learning methods
Prokofiev	composers		

Figure 2. *Analysis of text elements*

Aslib, The Association for Information Management
Information House, 20-24 Old Street, London EC1V 9AP
Tel: +44 (0) 171 253 4488, Fax: +44 (0) 171 430 0514
Email: pubs@aslib.co.uk, WWW: http://www.aslib.co.uk/aslib

Journal of Documentation, Vol. 53, No. 2, March 1997

© Aslib, The Association for Information Management.

All rights reserved. Except as otherwise permitted under the Copyright, Designs and Patents Act 1988, no part of this publication may be reproduced, stored in a retrieval system, or transmitted in any form or by any means, electronic, mechanical, photocopying or otherwise without the prior written permission of the publisher.

wood', 'made of skin', 'frets', 'Africa', 'New Orleans', 'Europe', 'fixed by screws', 'flat sides', 'back of the body', 'played with the thumb', 'different sizes' and 'sweet pitch', that are mentioned once only in the four definitions for banjo.

The categories identified above are based on the meaning of the relevant text elements in the definitions. The term banjo will be taken to exemplify the selection process of the names of the categories. The text element 'instrument' is a *kind of sound device*, thus the category *kind of sound device/solo* is assigned to it. 'Stringed', related to musical instruments, refers to the source of sound or to where the sound comes from, so this is the category identified for it. A 'plucked' musical instrument is one that produces the sound by plucking it, thus plucked refers to the *manner of producing sound*, and this is the category created for it. 'With the fingers' alludes to the means by which the sound is produced, so the *means of producing sound* category is suggested. 'Neck, strings, body, frets, sides' etc. refer to the physical/morphological description of the instrument being defined; thus the *morphological/physical* category is identified. 'Wood' and 'skin' refer to the material the instrument is made out of, for them the *material* category is suggested. The same procedure has been followed with the relevant text elements identified for the other nine musical instruments.

Based on the categories for the ten musical instruments already identified, a unified list of these categories can be suggested. This list will be the real departure point for building the structure (see Figure 3).

In the categories identified in Figure 3, there are two main groups. One of them includes the categories that represent the solo instruments themselves: materials, morphological characteristics, type of sound device/solo, source of sound, manner of producing sound, means of producing sound, notation, significance-role, musical character, musical functions, playing techniques, status, relation with other instruments and social functions. The other one includes categories not directly related to the solo instruments. In this group two different branches are identified: one is music related, and is represented by categories such as musical genre, musical ensembles, orchestras, composers, compositions, instrument makers, music teaching centres, works on music, performers, learning methods and theoreticians that have to be located in the general structure for the field of music. The other new group is not music related but use related as it deals with ethnic characteristics, time, military/religious groups and place. This latter new group needs also to be represented in the thesaurus structure. This means that, from the definitions themselves, we have information that allows structuring the musical instruments group. Moreover, we have information that allows identifying other music categories closely related to the musical instruments. Furthermore, we have information at hand to distinguish relevant categories dealing with other subject fields – different from music – which are also relevant for the sample case. This means that we can elaborate a complete set of conceptual relationships for the discipline of music, including those fields not music related, from the information found in the dictionary definitions. This information can be later represented in the thesaurus structure.

The suggested categories in Figure 3 are based on relevant aspects of the

Aslib, The Association for Information Management
Information House, 20–24 Old Street, London EC1V 9AP
Tel: +44 (0) 171 253 4488, Fax: +44 (0) 171 430 0514
Email: pubs@aslib.co.uk, WWW: http://www.aslib.co.uk/aslib

Journal of Documentation, Vol. 53, No. 2, March 1997

© Aslib, The Association for Information Management.

All rights reserved. Except as otherwise permitted under the Copyright, Designs and Patents Act 1988, no part of this publication may be reproduced, stored in a retrieval system, or transmitted in any form or by any means, electronic, mechanical, photocopying or otherwise without the prior written permission of the publisher.

1. type of sound device
2. source of the sound
3. manner of producing sound
4. means of producing sound
5. physical/morphological characteristics
6. material
7. musical character
8. notation
9. playing techniques
10. musical functions/*
11. musical functions/forms
12. musical functions/genre
13. musical function/ensembles
14. musical function/orchestras
15. social functions
16. status
17. relation with other instruments/*
18. relation with other instruments/ family
19. relation with other instruments/ origins
20. relation with other instruments/ influences
21. ethnic characteristics
22. religious/military groups
23. place
24. time
25. significance-role
26. composers
27. compositions
28. instrument makers
29. performers
30. music teaching centres
31. theoreticians
32. works on music/titles
33. learning methods
34. musical functions/arrangements
35. musical functions/scores

Figure 3. *Unified list of categories*

*to be named when more data at hand

musical instruments. This is one important step in getting the generators-texts structure into that of the thesaurus. It is also interesting to note that categories such as physical/morphological characteristics are rarely found in actual classifications of instruments, as a main criterion for subdivision, although they might be found as distinguishable criteria down in the hierarchy. The same thing can be said regarding the social and musical functions categories. The fact of identifying 'unusual' categories after following the above method might be seen as of not much importance by itself, but we think it is, because this means that the proposed method has the capability of recovering 'hidden' relevant aspects for a given concept that, otherwise, might be unseen or underestimated.

Regarding the validity of the suggested categories, we must remind ourselves that they come directly from the analysis of the statements found in dictionaries and encyclopedias of music for single concept terms. It is interesting to note that some of the categories are also present in some musical instrument classifications based on oral and/or non-Western traditions, collected by Kartomi [39]. For example, the Javanese classification of musical instruments, which has a primarily oral orientation, has the instruments' basic material, and their size and shape, as general criteria for classification. For this reason, we think that our method leads to a recognition of relevant criteria for organising the musical instruments group, and that these criteria are also found relevant by people. This means that the proposed structure would include aspects or elements and organising criteria that seem to be close to the user's model in this subject field. This fact makes us think that this method can be of help in carrying out our initial goal of building a more interactive thesaurus to improve communication in IR systems.

Aslib, The Association for Information Management
Information House, 20-24 Old Street, London EC1V 9AP
Tel: +44 (0) 171 253 4488. Fax: +44 (0) 171 430 0514
Email: pubs@aslib.co.uk. WWW: http://www.aslib.co.uk/aslib

Journal of Documentation, Vol. 53, No. 2, March 1997

© Aslib, The Association for Information Management.

All rights reserved. Except as otherwise permitted under the Copyright, Designs and Patents Act 1988, no part of this publication may be reproduced, stored in a retrieval system, or transmitted in any form or by any means, electronic, mechanical, photocopying or otherwise without the prior written permission of the publisher.

Summarising, we can say that the proposed method allows us:

(a) to identify a large number of categories under which to structure the thesaurus vocabulary – due to its capacity of discovering 'hidden' categories in the definitions. As a consequence, each concept included in the thesaurus will be more and better represented in the thesaurus structure;
(b) to represent each concept in multiple ways, according to the different nature of these relevant categories;
(c) to validate the suggested categories;
(d) to rank the importance of the categories identified, according to the authors' information, as will be demonstrated below.

3.2 Organising the categories identified in the structure

The selection of relevant categories based on the relevant text elements found in the definitions is an important part in the process of building the thesaurus structure. It could be considered as the most important one. The next feature that needs to be addressed is the organisation of these categories in the structure, and the conceptual relations among them. This is also an important step. In our view, this organisation should reflect the categories' frequency of citation in the definitions and the generator preferences as far as possible. Some preferences can be detected, for example, by the way they structure the text. Some authors start the definition by mentioning the instrument's ethnic aspects, but others do so by discussing musical aspects or physical characteristics. Subjective information in the definitions is represented by the category 'instruments' significance-role' in our sample that stands for text elements such as 'main instrument, essential in the orchestras, known in the whole world', and the like.

The way a category might be broken down, according to the authors' knowledge and preferences, can be seen as in close relation with cognitive processes. We believe that the structure development based on the categories' frequency of citation in the definitions and on the generators' preferences, might help in increasing the thesaurus interaction capabilities.

We have recorded the frequency of citation of the unified categories, seen in Figure 3, and this ranking has been established for the ten musical instruments in our sample. This statistical account has been graphically represented in Figure 4.

This table shows the selected categories represented by numbers, ranging from one to thirty-five as they are numbered in Figure 3. The number of times a given category is mentioned in the definitions for a term has been recorded. The letters A, B, C and D inside the cells refer to the four dictionaries used in this study. The highest score for a category to have is four if we talk about a single instrument. If we refer to the whole group the highest score for a category to have is forty. Some letters carry an asterisk at their right side. This means that they are inferred from the definition of that particular dictionary. It is important to establish this distinction, because the categories expressed in the

Aslib, The Association for Information Management
Information House, 20–24 Old Street, London EC1V 9AP
Tel: +44 (0) 171 253 4488, Fax: +44 (0) 171 430 0514
Email: pubs@aslib.co.uk, WWW: http://www.aslib.co.uk/aslib

Journal of Documentation, Vol. 53, No. 2, March 1997

© Aslib, The Association for Information Management.

All rights reserved. Except as otherwise permitted under the Copyright, Designs and Patents Act 1988, no part of this publication may be reproduced, stored in a retrieval system, or transmitted in any form or by any means, electronic, mechanical, photocopying or otherwise without the prior written permission of the publisher.

| Instrument | Categories | | |

Journal of Documentation, Vol. 53, No. 2, March 1997

© Aslib, The Association for Information Management.

All rights reserved. Except as otherwise permitted under the Copyright, Designs and Patents Act 1988, no part of this publication may be reproduced, stored in a retrieval system, or transmitted in any form or by any means, electronic, mechanical, photocopying or otherwise without the prior written permission of the publisher.

definitions have a bigger weight than the inferred ones. This distinction weights the final ranking.

We understand that even though the final scores for two categories are the same, we state that they are not equally relevant if one of them has no inferences, that is, one has A, B, C and D in its cell, while the other has A*, B*, C* or D*. The asterisk introduces a qualitative criterion that is specially interesting and necessary when two or more categories have the same final ranking, because it allows their relevance to be weighted.

Taking balalaika as an example, we see that is has ABCD in cell number 5, which stands for physical/morphological characteristics. This means that this category is highly relevant for the balalaika. If all the instruments have ABCD in cell number 5, it means that this category is highly relevant for the instruments group. Likewise, we see in Figure 4 that categories 1 and 2 yield the same final scores – forty – but 2 has some inferred categories, and 1 has none. This means that 2 is less relevant than 1, and that it would come after it in the thesaurus structure, as we shall see later.

We have represented the frequency of citation in a different graphic in Figure 5, where the same data that appear in Figure 4 are expressed in a different way. Figure 5 allows checking the importance or the relevance level of the categories listed in Figure 3.

After having identified the relevant categories for the group of musical instruments, and after knowing their frequency or their relevance level, we can start developing the thesaurus structure. Going from top to bottom, the categories to be cited first in the structure, according to Figure 4, should be type of sound device/solo, morphological/physical characteristics, source of sound, manner of producing sound and means of producing sound. With less representation, we find the categories related to musical and social functions. None of them alone is mentioned in every instrument, but if we consider instead that functions encompass all, we see that this category is mentioned in all the instruments. So, we may consider functions also at the top level.

The top categories need to be organised according to their similarities, because it is obvious that they belong to different classes. When this is done, the group with more than one category will be broken down according to the data in Figure 4 to establish a degree of relevance. For instance, in the group that could be named 'related to sound', the category type of sound device/solo would come first in the structure because it is cited four times in each instrument; source of sound would come after it, because some of its citations are inferred. Down in the structure we should place manner of producing sound and means of producing sound. The functions group seems easy to develop, since it has two well defined categories, musical functions and social functions.

The suggested structure for the categories most frequently cited would be as follows:

1. Categories related to sound
   type of sound device/solo

Aslib, The Association for Information Management
Information House, 20–24 Old Street, London EC1V 9AP
Tel: +44 (0) 171 253 4488, Fax: +44 (0) 171 430 0514
Email: pubs@aslib.co.uk, WWW: http://www.aslib.co.uk/aslib

Journal of Documentation, Vol. 53, No. 2, March 1997

© Aslib, The Association for Information Management.

All rights reserved. Except as otherwise permitted under the Copyright, Designs and Patents Act 1988, no part of this publication may be reproduced, stored in a retrieval system, or transmitted in any form or by any means, electronic, mechanical, photocopying or otherwise without the prior written permission of the publisher.

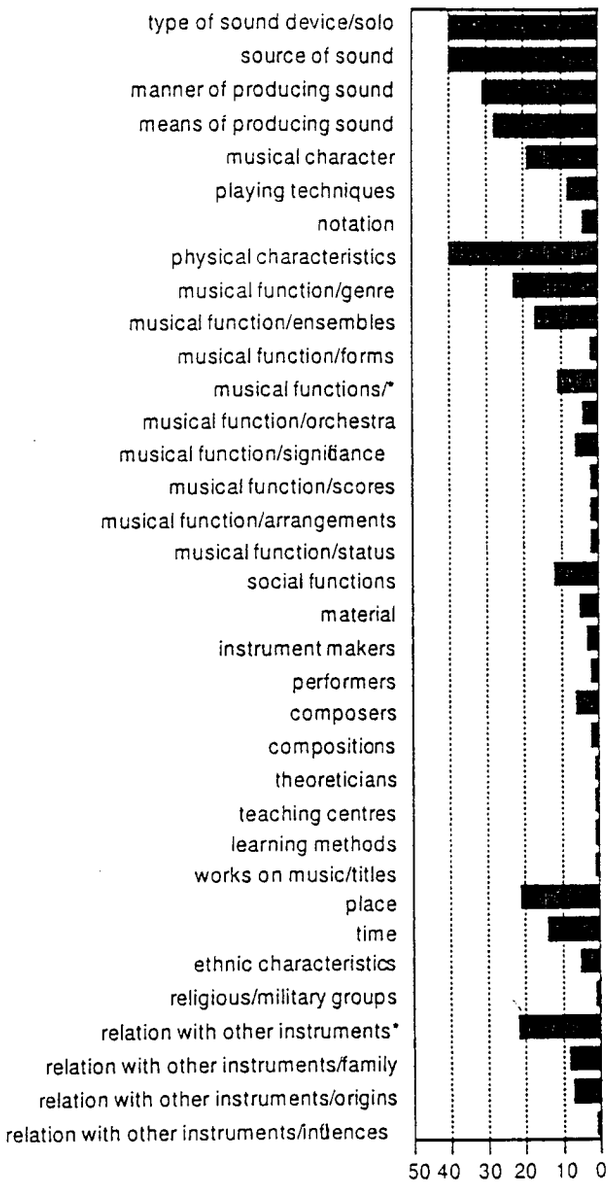


Figure 5. *Frequency of citation/relevance level of categories*

source of sound
manner of producing sound
means of producing sound

2. Categories related to functions
functions
musical functions

Aslib, The Association for Information Management
Information House, 20-24 Old Street, London EC1V 9AP
Tel: +44 (0) 171 253 4488, Fax: +44 (0) 171 430 0514
Email: pubs@aslib.co.uk, WWW: http://www.aslib.co.uk/aslib

Journal of Documentation, Vol. 53, No. 2, March 1997

© Aslib, The Association for Information Management.

All rights reserved. Except as otherwise permitted under the Copyright, Designs and Patents Act 1988, no part of this publication may be reproduced, stored in a retrieval system, or transmitted in any form or by any means, electronic, mechanical, photocopying or otherwise without the prior written permission of the publisher.

social functions

3. Categories related to morpho-physical description

Once the main groups are arranged, and the most representative categories are organised within them, the remaining categories with lower representation in the definitions shall be located within the structure. The criterion to be followed in doing so is the same that has been mentioned previously: the ranking of categories shown in Figure 4. According to this, the group 'categories related to sound' has to be completed with the remaining 'sound' categories in the following order: musical character, playing techniques and notation. The 'related to sound' group needs further development. Two upper categories are required, because they are implicit in this group. Type of sound device/solo, reminds us that other types of sound devices might be encountered, so we need a category to encompass them all, such as type of sound devices/instruments. Still, an upper category for all types of sound devices that may exist would be necessary, such as sound devices. The functions group, already divided into two main branches – musical functions and social functions – would be developed as follows. Under musical functions, we should locate musical functions/genre, musical functions/forms, musical functions/ ensembles, musical functions/orchestras, musical functions/* instruments significance and status. Under social function we would include popular instruments.

The 'material' category (see Figure 3) deserves special attention. It has not much representation in the definitions, according to the data in Figure 4, which means that it is a secondary criterion for structuring the sample group of instruments. Nevertheless, it is expected that it gains representation when all the terminology, regarding the musical instruments group, is processed. Meanwhile, we shall include it in the structure within brackets to mean that it is a hypothetical position. This category may be organised according to the kinds of materials the instruments are made of.

It will be necessary to locate in the general structure the following cases: (a) some categories related to the function group, such as musical genre, musical forms, scores, ensembles and orchestras; (b) other music categories, such as instrument making, learning methods, teaching centres, composers, performers, instrument makers, theoreticians and sources/information on music; and (c) other categories that do not belong to the field of music, such as place, time, ethnic characteristics and religious/military groups.

Finally, it is interesting to discuss the 'relation with other instruments' category. The identification of specific semantic relations that may exist among the different instruments is not the goal of the present work and does not matter much at this stage. Nevertheless, this category links a given instrument with others. It is shared by expressions broader than the solo instrument, as is the case of the violin family in the violin example, and by associative expressions related to the solo instrument, such as is the case of lute in the mandolin example. So, it is quite beyond the standard rules. That is why it is not located in the structure for the moment. Nevertheless, we find it very expressive, because it gives a sketch of a given instrument at a glance. Origins, instrument family and

Aslib, The Association for Information Management
Information House, 20–24 Old Street, London EC1V 9AP
Tel: +44 (0) 171 253 4488, Fax: +44 (0) 171 430 0514
Email: pubs@aslib.co.uk, WWW: http://www.aslib.co.uk/aslib

Journal of Documentation, Vol. 53, No. 2, March 1997

© Aslib, The Association for Information Management.

All rights reserved. Except as otherwise permitted under the Copyright, Designs and Patents Act 1988, no part of this publication may be reproduced, stored in a retrieval system, or transmitted in any form or by any means, electronic, mechanical, photocopying or otherwise without the prior written permission of the publisher.

the influences on other instruments, which are the subcategories, illustrate this brief history. Besides, if we think of the users, we believe that 'relation with other instruments' is closer and easier for them to use than expressions such as 'generic term' or 'related term' that might not mean much to them.

The complete suggested structure for the musical instrument, according to the definitions, as well as other general categories required in the music thesaurus, also based on the definitions is presented in Figure 6.

An interesting thing about the method that we have already described is that there is no need either for searching for the 'right' category or for the 'right' ordering of categories in the structure. The structure is simply modelled on the texts, and because of that, according to their authors' knowledge and intentions.

3.3 Concept representation

The selection of categories according to what is relevant for the authors of the dictionaries, and the different nature of these categories, make possible the polyrepresentation of concepts. Polyrepresentation means in this context the chances for a concept to be represented in as many categories as are relevant for it, according to the definitions. It also means that there will be an exhaustive representation for each concept in the thesaurus structure. It has to be noted that it has nothing to do with the polyhierarchy. We think that this feature will facilitate the user searching process, for the simple reason that if one concept is included in six or seven locations in the structure, there are more chances for it to be found than if it is in only one. The principle of variety [1, pp. 361–363] and the principle of redundancy [1, 18] – we would say in this case 'controlled redundancy' – are somehow present in the structural polyrepresentation of concepts. At the same time, this feature will help in the 'hit side of the barn' principle [40]. As a result, an increase of recall in information retrieval should be expected.

To give a quick illustration of what we mean, we would say that any musical instrument will have at least a structural representation under source of sound, physical characteristics, functions and material; these being only the most representative categories.

Going down in each category the polyrepresentation increases. A given instrument, let us say violoncello, could be found under musical functions/* because 'it is a soloist instrument'; under musical functions/forms, because 'it plays sonatas, concertos and ricercares'; under musical functions/genre, because 'it plays instrumental, dramatic and chamber music'; and under musical functions/orchestras, because 'it is played in the orchestra'. If we look at the sound categories, the violoncello can be represented under source of sound, because 'it is a stringed instrument'; under manner of producing sound, because 'its strings are scraped by a bow'; under means of producing sound, because it is a 'bowed instrument'; under musical character, because 'it has a deep sonority' and 'it is tuned doh, sol, re, la'; and, finally, under playing techniques, because 'it is held between the knees'. The same will happen in the case of the morphological/physical characteristics category when it is developed. The

Aslib, The Association for Information Management
Information House, 20–24 Old Street, London EC1V 9AP
Tel: +44 (0) 171 253 4488, Fax: +44 (0) 171 430 0514
Email: pubs@aslib.co.uk, WWW: http://www.aslib.co.uk/aslib

Journal of Documentation, Vol. 53, No. 2, March 1997

© Aslib, The Association for Information Management.

All rights reserved. Except as otherwise permitted under the Copyright, Designs and Patents Act 1988, no part of this publication may be reproduced, stored in a retrieval system, or transmitted in any form or by any means, electronic, mechanical, photocopying or otherwise without the prior written permission of the publisher.

sound devices
  sound devices/instruments
    type of sound devices/solo
      source of sound
        manner of producing sound
          means of producing sound
            musical character
              playing techniques
                notation
    functions
      musical functions
        musical functions/genre+
          musical functions/arrangements+
          musical functions/scores+
        musical functions/forms+
          musical functions/arrangements+
          musical functions/scores+
        musical functions/ensembles+
        musical functions/orchestras+
        musical functions/*
        musical functions/significance
        musical functions/status
      social functions
        popular instruments
          place+
          time+
          ethnic characteristics+
          religious/military groups+
    [material]
    type of sound devices/ensembles
    type of sound devices/orchestras
musical genre
musical forms
instrument making
teaching of music
  learning methods
  teaching centres
  composition
sources/information on music
  works on music
persons/names
  composers
  performers
  theoreticians
  instrument makers
ethnic characteristics
religious/military groups
place
time

Figure 6. *Suggested structure for musical instruments*

+ indicates categories which show the relationship between the sound device and other categories in the structure. They are to be developed in the other places in the structure to which they belong.

Aslib, The Association for Information Management
Information House, 20–24 Old Street, London EC1V 9AP
Tel: +44 (0) 171 253 4488, Fax: +44 (0) 171 430 0514
Email: pubs@aslib.co.uk, WWW: http://www.aslib.co.uk/aslib

Journal of Documentation, Vol. 53, No. 2, March 1997

© Aslib, The Association for Information Management.

All rights reserved. Except as otherwise permitted under the Copyright, Designs and Patents Act 1988, no part of this publication may be reproduced, stored in a retrieval system, or transmitted in any form or by any means, electronic, mechanical, photocopying or otherwise without the prior written permission of the publisher.

polyrepresentation of concepts is possible, because we have provided a suitable structure for this to occur.

The solo instruments group has a semantic relationship with other musical categories, such as musical ensembles and musical genre, and other non music categories, such as the ethnic groups the instrument is most commonly linked to, or the places it is from, or the time when it was used, to mention some of them. This linking is made by connecting the appropriate category within the solo instruments, with the other categories already identified in the thesaurus structure, which stand for places, ethnic groups, musical genres, type of sound devices/ensembles, musical forms etc. (see Figure 6). The category function seems to be the one most likely for housing most of these links. For instance, under social functions, 'popular instruments' is located. This concept is closely related to places, ethnic groups and religious/military groups, as we can see in the definitions. This link is pointed out in the structure by writing the place, the ethnic characteristics and the religious/military groups under popular instruments (see Figure 6).

Polyrepresentation also means that if a place is provided in the structure for each relevant text element found in the definitions, it would make it possible that a document about any instrument can be indexed under a particular place: thus Africa, Europe, USA, New Orleans etc. → places category; a particular ethnic group: black people, Arab people etc. → ethnic or similar category; a particular musical genre: instrumental music, dramatic music, domestic music, popular music etc. → musical genre; a particular ensemble/orchestra: jazz bands, popular ensembles → musical ensembles, and under the instruments' name themselves; banjos, violoncellos etc. → source of sound, physical characteristics, material, and functions in the sound devices category.

The proposed method for building the thesaurus structure allows us to work at two levels: a macrostructural level, which is represented by the overall main categories for the subject domain we are interested in, whether they are music related or not. The second level is a microstructural one, which is represented by the categories identified as relevant in the definitions for a given concept, in our example the musical instruments group. The microstructural level seems to be of special interest, as a way of materialising the concept of poly-representation in the structure, and as a means of identifying new or 'unexpected' conceptual relations for the concept. An example of this can be recognised in the fact that material is a category for the instruments to be represented in, although they are not a material, morphological characteristics is another suitable category for the musical instruments, although they are not morphological characteristics, and functions is another convenient category for the instruments to be in, although they are not functions if we look at them from a macrostructural point of view. This can lead to the identification of new semantic relations for the musical instruments that we may want to be shown in the record for each of them. These new relations can be of the following kind: coming from, made out of, genre played with, musical forms played with, what it looks like, linked ethnic groups, compositions for it, related composers, learning methods etc. These relations fall into the concept of what we usually

Aslib, The Association for Information Management
Information House, 20–24 Old Street, London EC1V 9AP
Tel: +44 (0) 171 253 4488, Fax: +44 (0) 171 430 0514
Email: pubs@aslib.co.uk, WWW: http://www.aslib.co.uk/aslib

Journal of Documentation, Vol. 53, No. 2, March 1997

© Aslib, The Association for Information Management.

All rights reserved. Except as otherwise permitted under the Copyright, Designs and Patents Act 1988, no part of this publication may be reproduced, stored in a retrieval system, or transmitted in any form or by any means, electronic, mechanical, photocopying or otherwise without the prior written permission of the publisher.

call related terms. We think that if the thesaurus identifies specific relations, instead of listing descriptors under the name related terms in the record for each concept, it would be more helpful to the user. The master record for the musical instruments would have the traditional generic-specific relations and the relations newly identified that would take the place of the traditional related terms. It would include information of the following kind, regarding the specific semantic relationships identified in the definitions:

Instrument name

musical character
coming from/place related to
time
made out of
musical genre played with
musical forms played with
musical ensembles related to
orchestras related to
relation with other instruments
what it looks like/morphological characteristics
ethnic groups linked to
religious/miliatry groups linked to
related composers
compositions for it
learning methods

As a result of the possibilities for polyrepresentation of a concept in the structure and of the identification of new semantic relations for inclusion in the record of the descriptor, we can predict an increase of recall in information retrieval. This is an important feature for an information retrieval tool, since it is known that both conventional and non-conventional IR systems are reporting low recall rates in their performances.

Lancaster *et al.* refer to a low recall in a conventional IR system. They found that many relevant records were not retrieved even in very broad searches. They estimated an average recall of 53.9% [41]. In non-conventional IR systems the situation is worse. Blair evaluated recall and precision in a full-text retrieval system. The results yield an average recall of 20% and a precision of 79.0% [42]. Maron also noted the poor performance of IR systems. He suggested that recall deteriorates as documents in the system grow in size [43, 44].

Many have argued that the thesaurus structure, and the semantic relationships in the thesaurus, are considered recall oriented devices [45]. If we are able to create a rich structure for a given subject, that is one that allows for the polyrepresentation concepts, and if we can enrich the traditional semantic relationships that appear nowadays in the thesauri, as is shown above, we would say that we are improving the recall performance of actual thesauri.

4. USERS AS A SOURCE FOR THE THESAURUS STRUCTURE

The other pillar, and the most challenging one, for constructing the thesaurus structure, is the user. This part of the paper is needed to understand the previous part in its context which is why it is included here, even though no practical research has been done yet. As it is based mainly on theoretical

Aslib, The Association for Information Management
Information House, 20–24 Old Street, London EC1V 9AP
Tel: +44 (0) 171 253 4488, Fax: +44 (0) 171 430 0514
Email: pubs@aslib.co.uk, WWW: http://www.aslib.co.uk/aslib

Journal of Documentation, Vol. 53, No. 2, March 1997

© Aslib, The Association for Information Management.

All rights reserved. Except as otherwise permitted under the Copyright, Designs and Patents Act 1988, no part of this publication may be reproduced, stored in a retrieval system, or transmitted in any form or by any means, electronic, mechanical, photocopying or otherwise without the prior written permission of the publisher.

assumptions, no results or examples can be given. Nevertheless, it will outline the problem, and present some tentative methods for solving it.

Shaping what is relevant for the user in a particular subject domain is a complex matter, but almost everybody agrees that topicality is not the only aspect constituting relevance [46–48]. If we want the structure to respond to the users' information needs, and if we intend to articulate these needs in a familiar way for the user, there is no doubt that knowledge, intentionality, own purpose, and, in general, the elements indicated by Ingwersen constituting the user's cognitive space [26, p. 103] have to be taken into consideration as far as possible. This thinking is also very much stressed by Solomon in an interesting work describing methods for user-based classification development [49]. To simplify this complexity, two different parts are recognised when analysing the user's information needs. One of them deals with his or her actual knowledge of the domain, and with his or her model of the subject domain, and the other deals with his or her goals, special tasks, jobs and preferences etc., on the understanding that the addition of two parts would result in a satisfactory representation of the user's information needs. In other words, the addition of both parts would give us the usefulness of the information retrieved. After identification of these relevant elements, the analysis of the former – user knowledge – would generate relevant elements for representing the subject domain from the user's point of view which may lead to relevant categories for the field. The analysis of the latter – the user's own purpose – would yield relevant elements not related to the topic covered by the thesaurus, but equally interesting and important. From the non-topical elements, new categories may be identified which would be included in the thesaurus structure as well. Finally, the results obtained would be checked against the structure generated by the authors. One merging structure, joining categories from authors' texts and users, may be built at the end of the process.

The identification of relevant elements, from the user's point of view, that later would be located in the thesaurus structure will be our concern in this stage. The written text of the user's information needs could be one of the sources to be considered. There are two main reasons for that. Firstly, recent studies stated that the users' written question statements are the most effective source for analysing information needs [50]. We would define the user's information needs as the subject's written statements of his or her information needs at the different stages of the information seeking process he or she is going through, because of his or her anomalous state of knowledge [51, 52]. Secondly, it is possible that the same methodology described in part 3 for identifying relevant elements in the generators' writings can be used in analysing the user's written text.

Relevance in IR systems is not static, but dynamic. Relevance judgements may change during the information retrieval process until the user's knowledge gap has been filled. In other words, information seeking is a dynamic process where the user's context can be modified by relevant findings [47] thus relevance judgements are expected to change during the process. For this reason, the written query statement is expected to change in the process, and each time this

Aslib, The Association for Information Management
Information House, 20–24 Old Street, London EC1V 9AP
Tel: +44 (0) 171 253 4488, Fax: +44 (0) 171 430 0514
Email: pubs@aslib.co.uk, WWW: http://www.aslib.co.uk/aslib

Journal of Documentation, Vol. 53, No. 2, March 1997

© Aslib, The Association for Information Management.

All rights reserved. Except as otherwise permitted under the Copyright, Designs and Patents Act 1988, no part of this publication may be reproduced, stored in a retrieval system, or transmitted in any form or by any means, electronic, mechanical, photocopying or otherwise without the prior written permission of the publisher.

happens a new written query statement from the user would be needed. So, analysing the written text of the user's information needs statement, and considering this statement as being dynamic, though subjected to changes, until the information retrieval process is over, would be one of the methods to be used.

The analysis of the query statements is not the only way of identifying relevant elements, topical and non-topical, from the users' point of view. Written texts themselves are also an interesting source for that. In our case, these texts are supposed to be specialised in a particular discipline or subject field, in our sample the field of music. The study would focus on what makes texts on music relevant for its potential users. This method can be considered as complementary to the study of the users' written query statements, and it will probably enrich the data obtained from the previous method. There is empirical evidence that confirms that there are characteristics other than topical relevance which affect document usefulness, or document relevance from a person's point of view. This topical relevance can be split into its constituent elements [53]. This study by Cool *et al.* presents results from two different but complementary projects. Each of them was addressed to two distinct groups of people: undergraduates and humanities scholars. From the former, six general categories were found containing relevant characteristics in texts. Four out of the six were not topic related: format, presentation, values, and oneself. From the latter group, it was also clear that factors other than topical relevance were important when deciding whether or not to use a document for a particular information need, and that the oneself facet was strongly reinforced in the scholar's group. The oneself facet was found to be related to the goals and the problems that the user is trying to solve.

This research strongly supports what it has been said in this paper. To the question the authors ask about the possibilities of representation of such characteristics for information retrieval, we would say that it is indeed possible and desirable, and that this is what we are trying to arrive at when this research, now starting, is finished. Nevertheless, we have to point out that not all the non-topic characteristics might be suitable to be represented in a thesaurus structure, because there are other mechanisms already functioning in the information systems, such as descriptive cataloguing, that can take care of them. We would say that the non-topic relevant characteristics, dealing with the formal description of the document, can be represented by descriptive cataloguing. The rest of them can be represented in the thesaurus structure, as has been said.

Looking a step further in the development of the thesaurus structure, a technique used by Pejtersen [12] might be worthy of consideration. She wished to create a semantic frame for the thesaurus terminology. To do so, she used the mental association technique, a method taken from the psychology field. She showed a word at a time to the subjects, and they wrote the first words that came to their minds without thinking. This method would show the relations among the thesaurus terminology according to the people's way of thinking. It was also envisaged as a way of representing the user's tacit knowledge. It is a

Aslib, The Association for Information Management
Information House, 20–24 Old Street, London EC1V 9AP
Tel: +44 (0) 171 253 4488, Fax: +44 (0) 171 430 0514
Email: pubs@aslib.co.uk, WWW: http://www.aslib.co.uk/aslib

Journal of Documentation, Vol. 53, No. 2, March 1997

© Aslib, The Association for Information Management.

All rights reserved. Except as otherwise permitted under the Copyright, Designs and Patents Act 1988, no part of this publication may be reproduced, stored in a retrieval system, or transmitted in any form or by any means, electronic, mechanical, photocopying or otherwise without the prior written permission of the publisher.

very interesting approach, but more research has to be done in order to find out if this technique is compatible with a long stock of terminology – thousands; Pejtersen worked only with 223 terms. We need to know if a long list of unstructured semantically related terms would be of help to users. For a thesaurus with twenty thousand descriptors, the list of associative relations for each descriptor might have hundreds of terms listed without any discrimination. We are not sure about the usefulness of such a device in this context. The same author used a method in her classification of fiction that is also an interesting way of getting information from users for building the thesaurus structure [11].

More research is needed before starting the thesaurus structure construction based on the user's point of view. Hopefully, this research will contribute the data needed for the identification of common patterns in users' information needs, to be translated later to the thesaurus structure. The structure will be built according to this information.

5. CONCLUSIONS

A different conceptual approach for designing the structure of a subject domain oriented thesaurus has been attempted in this paper. It is thought of as a way of easing the communication between texts and users in order to increase the interaction in IR, and thus to facilitate information transfer. Kim talks about some attempts made to exploit a thesaurus to improve retrieval effectiveness [54], but no-one addresses the problem as is done here. It has also been said in this paper that improving retrieval effectiveness might be possible by creating a structure modelled on the generators and users as far as possible. It is our belief that by capturing the originators' and users' models in the structure, the information retrieval process would be shortened, because the user would identify the relevant documents among many of them quickly. This would be possible if the documents were clearly represented in the system, according to the criteria developed in this article. We think that if readers and authors communicate well through the writing when the reader's model and the author's model meet in that writing, something similar may occur in information retrieval, when the system provides a conceptual structure for this to happen. We would say that the conceptual structure described in this work facilitates this meeting. Figure 7 shows the thesaurus structure design model that has been developed in this paper.

The concept of subject domain in this context goes beyond the traditional one. Subject domain is understood here as an ensemble of knowledge plus other features such as tasks, jobs, intentionality and preferences which are part of the user's cognitive space, and that may greatly affect relevance in IR systems. This concept will modify the whole structural designing process, and the building itself, as could be seen previously. As a consequence, the structure in general, and the potential categories within it in particular, are not only 'subject' oriented, but also cognitively oriented. Following this thinking, two general kinds of categories are expected, categories representing the subject domain and categories representing what could be called here for simplicity, the generators' and the

Aslib, The Association for Information Management
Information House, 20–24 Old Street, London EC1V 9AP
Tel: +44 (0) 171 253 4488, Fax: +44 (0) 171 430 0514
Email: pubs@aslib.co.uk. WWW: http://www.aslib.co.uk/aslib

Journal of Documentation, Vol. 53, No. 2, March 1997

© Aslib, The Association for Information Management.

All rights reserved. Except as otherwise permitted under the Copyright, Designs and Patents Act 1988, no part of this publication may be reproduced, stored in a retrieval system, or transmitted in any form or by any means, electronic, mechanical, photocopying or otherwise without the prior written permission of the publisher.

users' own purpose. The concept of structure is thus much broader than in previous conceptions. As a consequence, it becomes richer and more powerful than in previous contributions.

Our approach sees authors, texts, users and system in a balanced relationship. We can represent it graphically in Figure 7.

The model shows the main features responsible for modelling texts: generators' knowledge and own purpose, and the main features modelling the users' information needs: knowledge and own purpose. The thesaurus structure in the system is the intermediary between texts and users. The thesaurus, called the intermediary structure, is the third main structure in IR systems, and it is designed according to the texts' and users' models, merging both of them in only one structure.

In theory, and according to the model, the user could encounter in the thesaurus structure the knowledge and purpose of the authors that he or she is looking for, and so remedy his or her lack of knowledge. All this will lead the user to relevant documents for a particular information need. It can be predicted thus that the results of information retrieval in such a system are expected to be satisfactory. Now, it is easy to understand that we do not consider the proposed thesaurus as an outside, convenient element in the information retrieval process, to complicate rather than to ease communication in IR systems. It is, on the contrary, one that is intended to be in close symbiosis with the two main elements in the system: texts and users. It is thought of as a device to represent the users' needs and the information contents of documents in equivalent ways.

As a result of this approach, new relationships among the concepts in the thesaurus structure, different from those already established, may appear. This will be motivated by our concept of subject domain, structure scope, and the method followed in putting the design principles into practice. Some of these new relations have already been mentioned (see p. 169).

We also believe that the structure might have a positive influence in interface

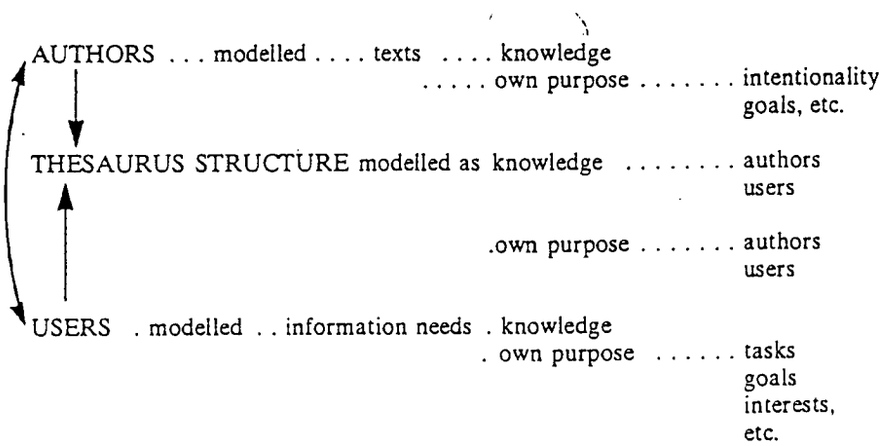


Figure 7. *Model for a thesaurus structure design*

Aslib, The Association for Information Management
Information House, 20–24 Old Street, London EC1V 9AP
Tel: +44 (0) 171 253 4488, Fax: +44 (0) 171 430 0514
Email: pubs@aslib.co.uk, WWW: http://www.aslib.co.uk/aslib

Journal of Documentation, Vol. 53, No. 2, March 1997

© Aslib, The Association for Information Management.

All rights reserved. Except as otherwise permitted under the Copyright, Designs and Patents Act 1988, no part of this publication may be reproduced, stored in a retrieval system, or transmitted in any form or by any means, electronic, mechanical, photocopying or otherwise without the prior written permission of the publisher.

design, because it is often based on semantic structures of different kinds [55, 56], and the structure described in this work is quite rich. All of this together leads to the possibility of a concept being polyrepresented in the structure.

Finally, an attempt has been made in this paper to describe the design of a thesaurus structure modelled according to the originators of information through their writings, and to the users' information needs. The presence of the user side in this paper is there mainly to contextualise the meaning of the generators' side and to give a general outline of what we intend to do. Further research is needed in order to allow the proposed structure to be built and tested. These first research results are intended to call attention to the possibility of developing the concept and the potential of the thesaurus and, mainly, of exploring further the thesaurus structure as a more suitable tool for use in IR systems.

ACKNOWLEDGEMENTS

The author acknowledges a grant from the Programa de Perfeccionamiento y Movilidad de Personal Investigador, DGICYT, Ministry of Education and Science, Spain, to carry out a research project at the School of Communication, Information and Library Studies, Rutgers University, New Brunswick, NJ, USA. She also acknowledges the reading and comments on the draft of this paper by Professors James Anderson and Nicholas Belkin at the School of Communication, Information and Library Studies, Rutgers University, New Brunswick, NJ, USA.

REFERENCES

1. Bates, M. Subject access in online catalogs: a design model. *Journal of the American Society for Information Science, 37*(6), 1986, 357–376.
2. Bates, M. Rethinking subject cataloging in the online environment. *Library Resources and Technical Services, 33*(4), 1989, 400–412.
3. Anderson, J.D. *and* Rowley, F.A. Builing an end-user thesaurus from full text. *In*: Kwasnik, B.H. *and* Fidel, R., *eds. Advances in classification research II. Proceedings of the 2nd ASIS SIG/CR: A classification Research Workshop*. Old Marlton Pike, NJ: Learned Information Inc., 1992, 1–13.
4. Chen, H., Yim, T. *and* Fye, D. Automatic thesaurus generation for an electronic community system. *Journal of the American Society for Information Science, 46*(3), 1995, 175–193.
5. Petersen, T. Developing a new thesaurus for art and architecture. *Library Trends, 38*(4), 1990, 644–658.
6. National Information Standards Organization. Proposed American national standard guidelines for construction, format, and management of monolingual thesauri: an American standard. Bethesda: NISO Press, 1994. (ANSI/NISO z39.19–1993).
7. Strong, G.W. *and* Drott, M.C. A thesaurus for end-user indexing and retrieval. *Information Processing and Management, 22*(6), 1986, 487–492.

Aslib, The Association for Information Management
Information House, 20–24 Old Street, London EC1V 9AP
Tel: +44 (0) 171 253 4488, Fax: +44 (0) 171 430 0514
Email: pubs@aslib.co.uk, WWW: http://www.aslib.co.uk/aslib

Journal of Documentation, Vol. 53, No. 2, March 1997

© Aslib, The Association for Information Management.

All rights reserved. Except as otherwise permitted under the Copyright, Designs and Patents Act 1988, no part of this publication may be reproduced, stored in a retrieval system, or transmitted in any form or by any means, electronic, mechanical, photocopying or otherwise without the prior written permission of the publisher.

8. Aitchison, J., Gomersall, R. *and* Ireland, R. *Thesaurofacet: a thesaurus and faceted classification for engineering and related subjects.* Whetstone: English Electric Company, 1969.
9. Bhattacharyya, G. Clasaurus: its fundamentals, design and use. *In*: Dahlberg, I., *ed. Universal classification and ordering systems.* Frankfurt: Indeks Verlag, 1982, 139–148.
10. Fugmann, R. An interactive clasaurus on PC. *International Classification, 17*(3/4), 1990, 133–137.
11. Pejtersen, A.M. Design of a classification scheme for fiction based on an analysis of actual user-librarian communications, and use of the scheme for control of the librarian's search strategies. *In*: Harbo, O. *and* Kajberg, L., *eds. International Research Forum on Information Science.* London: Mansell, 1980.
12. Pejtersen, A.M. *Interfaces based on associative semantics for browsing in information retrieval.* Roskilde, Denmark: Riso National Laboratory, 1991.
13. Foskett, D.J. Ranganathan and 'user-friendliness'. *Libri, 42*(3), 1992, 227–234.
14. Belkin, N. Cognitive models and information transfer. *Social Science Information Studies, 4*, 1984, 111–129.
15. Koll, M. *and* Srinivasan, P. Fuzzy versus probabilistics models for user relevance judgements. *Journal of the American Society for Information Science, 41*(4), 1990, 264–271.
16. Marchionini, G. Interfaces for end-user information seeking. *Journal of the American Society for Information Science, 43*(2), 1992, 156–163.
17. Kuhlthau, C.C. *Seeking meaning. A process approach to library and information services.* Norwood, NJ: Ablex Publishing Corporation, 1992.
18. Belkin, N. Anomalous state of knowledge for information retrieval. *Canadian Journal of Information Science, 5*, 1980, 133–143.
19. Waters, C. *and* Shepherd, M.A. Shifting the information paradigm from data-centered to user-centered. *Information Processing and Management, 30*(4), 1994, 455–471.
20. De Mey, M. The cognitive viewpoint: its development and its scope. *In*: De Mey, M., *ed. International Workshop on the Cognitive Viewpoint, 24–26 Mach 1977.* Ghent: University of Ghent, 1977, xvi–xxxii.
21. Ellis, D. The physical and cognitive paradigms in information retrieval research. *Journal of Documentation, 48*(1), 1992, 45–64.
22. Belkin, N. The cognitive viewpoint in information science. *Journal of Information Science, 16*, 1990, 1–15.
23. Harbo, O., Ingwersen, P. *and* Timmerman, P. Cognitive processes in information storage and retrieval. *In*: De Mey, M., *ed. International Workshop on the Cognitive Viewpoint, 24–26 March 1977.* Ghent: University of Ghent, 1977, 214–218.
24. Ingwersen, P. Search procedures in the library – analysed from the cognitive point of view. *Journal of Documentation, 38*(3), 1982, 165–191.
25. Belkin, N. Internal knowledge and external information. *In*: De Mey, M.,

Aslib, The Association for Information Management
Information House, 20–24 Old Street, London EC1V 9AP
Tel: +44 (0) 171 253 4488, Fax: +44 (0) 171 430 0514
Email: pubs@aslib.co.uk, WWW: http://www.aslib.co.uk/aslib

Journal of Documentation, Vol. 53, No. 2, March 1997

© Aslib, The Association for Information Management.

All rights reserved. Except as otherwise permitted under the Copyright, Designs and Patents Act 1988, no part of this publication may be reproduced, stored in a retrieval system, or transmitted in any form or by any means, electronic, mechanical, photocopying or otherwise without the prior written permission of the publisher.

ed. International Workshop on the Cognitive Viewpoint, 24–26 March 1977. Ghent: University of Ghent, 1977, 187–194.

26. Ingwersen, P. Elements of a cognitive theory for information retrieval interaction. *In*: Croft, B. *and* Van Rijsbergen, C.J., *eds. Proceedings of the 17th Annual International ACM-SIGIR Conference on Research and Development in Information Retrieval.* London, New York: Springer-Verlag, 1994.
27. Stiles, W.G. Ranganathan, cognition and expert systems. *Canadian Journal of Information Science, 10*, 1985, 16–24.
28. Brookes, B.C. The developing cognitive viewpoint in information science. *In*: De Mey, M., *ed. International Workshop on the Cognitive Viewpoint, 24–26 March 1977.* Ghent: University of Ghent, 1977, 195–203.
29. Ingwersen, P. *Information retrieval interaction.* London: Taylor Graham, 1992.
30. Flower, L. *and* Hayes, J.R. A cognitive process theory of writing. *College Composition and Communication, 32*, 1981, 365–387.
31. Flower, L. Taking thought: the role of conscious processing in making meaning. *In*: Maimon, E.P., Nodine, B.F. *and* O'Connor, F.W., *eds. Thinking, reasoning and writing.* New York: Longman, 1989.
32. Belkin, N. Using discourse analysis for the design of information retrieval mechanisms. *In*: Kuehn, J.J., *ed. Reseach and Development in Information Retrieval. Proceedings of the Sixth ACM SIGIR Conference.* New York: ACM, 1983, 31–47.
33. Budd, J. *and* Raber, W. Discourse analysis: method and application in the study of information. *Information Processing and Management, 32*(2), 1996, 217–226.
34. Van Dijk, T.A. *and* Kintsch, W. *Strategies of discourse comprehension.* New York: Academic Press, 1983.
35. Pérez, M. *Diccionario de la música y los músicos.* Madrid: Istmo, 1985.
36. Peña, J. *and* Anglés, H. *Diccionario de la música labor.* Barcelona: Labor, 1954.
37. Valls, M. *Diccionario de la música.* 4th. ed. Madrid: Alianza Editorial, 1982.
38. Sardá, A. *Léxico tecnológico musical en varios idiomas.* Madrid: Unión Musical Española, 1929.
39. Kartomi, M.J. *On concepts and classifications of musical instruments.* Chicago: University of Chicago Press, 1990.
40. Bates, M. The design of browsing and berrypicking techniques for the online search interface. *Online Review, 13*, 1989, 407–424.
41. Lancaster, F.W. *et al.* Identifying barriers to effective subject access in library catalogs. *Library Resources and Technical Services, 35*(4), 1991, 377–391.
42. Blair, D.C. Full text retrieval: evaluation and implications. *International Classification, 13*(1), 1986, 18–23.
43. Maron, M.E. Probabilistics design principles for conventional and full-text retrieval systems. *Information Processing and Management, 24*(3), 1988, 249–255.

Aslib, The Association for Information Management
Information House, 20–24 Old Street, London EC1V 9AP
Tel: +44 (0) 171 253 4488, Fax: +44 (0) 171 430 0514
Email: pubs@aslib.co.uk. WWW: http://www.aslib.co.uk/aslib

Journal of Documentation, Vol. 53, No. 2, March 1997

© Aslib, The Association for Information Management.

All rights reserved. Except as otherwise permitted under the Copyright, Designs and Patents Act 1988, no part of this publication may be reproduced, stored in a retrieval system, or transmitted in any form or by any means, electronic, mechanical, photocopying or otherwise without the prior written permission of the publisher.

44. Blair, C. *and* Maron, M.E. An evaluation of retrieval effectiveness for a full-text document-retrieval system. *Communications of ACM*, *28*(3), 1985, 289–299.
45. Gilchrist, A. *The thesaurus in retrieval*. London: Aslib, 1971.
46. Saracevic, T. Relevance: a review of a framework for the thinking on the notion in information science. *Journal of the American Society for Information Science, 26*, 1975, 321–343.
47. Harter, S.P. Psychological relevance and information science. *Journal of the American Society for Information Science, 43*(9), 1992, 602–615.
48. Schimber, L. *et al.* A re-examination of relevance: toward a dynamic situational definition. *Information Processing and Management, 26*(6), 1990, 755–776.
49. Solomon, P. User-based methods for classification development. *In*: Kwasnik, B.H. *and* Fidel, R., *eds. Advances in classification research II. Proceedings of the 2nd ASIS SIG/CR: A classification Research Workshop*. Old Marlton Pike, NJ: Learned Information Inc., 1992, 163–170.
50. Dalrymple, P.W. Retrieval by reformulation in two library catalogs: toward a cognitive model of searching behavior. *Journal of the American Society for Information Science, 41*(4), 1990, 272–281.
51. Belkin, N., Oddy, R. *and* Brooks, H.M. ASK for information retrieval. Part I. *Journal of Documentation, 38*, 1982, 61–71.
52. Belkin, N., Oddy, R. *and* Brooks, H.M. ASK for information retrieval. Part II. *Journal of Documentation, 38*, 1982, 147–164.
53. Cool, C., Belkin, N.J., Frieder, O. *and* Kantor, P. Characteristics of texts affecting relevance judgements. *In*: Williams, M.E., *ed. National Online Meeting. 14th, 1990, New York*. Medford, NJ: Learned Information, 1993, 77–84.
54. Kim, Y.W. *and* Kim, J.H. A model of knowledge based information retrieval with hierarchical concept graph. *Journal of Documentation, 46*(2), 1990, 113–136.
55. Jones, S. A thesaurus data model for an intelligent retrieval system. *Journal of Information Science, 19*, 1993, 167–178.
56. Jones, S., Gatford, M., Robertson, S., Hancock-Beaulieu, M., Secker, J. *and* Walker, S. Interactive thesaurus navigation: intelligence rules OK? *Journal of the American Society for Information Science, 46*(1), 1995, 52–59.

(*Revised version received 5 July 1996*)

Aslib, The Association for Information Management
Information House, 20–24 Old Street, London EC1V 9AP
Tel: +44 (0) 171 253 4488, Fax: +44 (0) 171 430 0514
Email: pubs@aslib.co.uk. WWW: http://www.aslib.co.uk/aslib

Synapse White Paper KMS101

Taxonomies, Ontologies, Thesauri and Authority Files: The Key to Better Information Retrieval

Copyright © 2003, by Synapse, the Knowledge Link Corporation

Contents

1. Hidden Information: The needle in the haystack
2. The Human Paradigm: How human beings solve the problem
3. The Machine Correlative: Syndetic networks - maps of knowledge
4. The Elasticity of language: Synonyms, homonyms and stretchable meanings
5. Hierarchical Perspectices: Tree structures, a top-down view
6. Associative Thinking: Making lateral connections
7. Electronic Thesauri: The ultimate vocabulary solution
8. Named Entity Control: Non-conceptual vocabularies
9. Personal Names
10. Organizational Names
11. Geographic Names
12. Enterprise Knowledge-Bases: Unifying knowledge throughout the enterprise
13. Six Degrees of Separation: Proximity in complex networks
14. Conclusions
15. References and Resources

Hidden Information

"Information networks straddle the world.
Nothing remains concealed. But the sheer
volume of information dissolves the information.
We are unable to take it all in."

Günther Grass

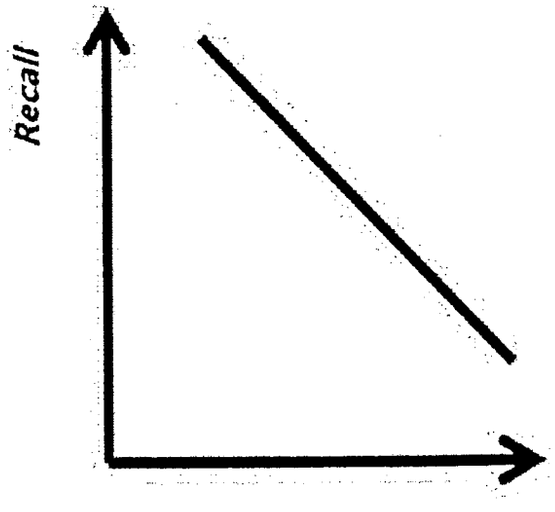
Few quotes so efficiently describe both the achievement and yet the failing of the information age at the beginning of the twenty-first century. We have terra-bytes of digitized information and global networks connecting millions of computers, but finding specific information is more than ever like looking for a needle in a haystack.

One of the major reasons why information retrieval is problematic is because language is elastic: one word can have many meanings, and one concept can be expressed by many different words. Human beings can understand words better than machines because our brains store a vast and complex web of interrelationships in which no word is an island.

Text-based searching has failed to provide a satisfactory method for interrogating large data sets and will fail increasingly as the sets of data we need to access continue their exponential growth.

The problems of text-based searching have been known in information science for decades as the precision-recall tradeoff.

Empirical studies have shown that precision will decline as recall increases and vice-versa; this tradeoff is inherent in the very nature of text-based searching.



Recall Quotient
relevant items retrieved / universe of relevant items
Precision Quotient
relevant items retrieved / total items returned

In other words, if you need to ensure you see all the relevant information then you will inevitably have to wade through a greater mass of irrelevant data – whatever you do to improve the comprehensiveness of your search will diminish the proportional purity of the results.

Back to Table of Contents

The Human Paradigm

"What a piece of work is man!
How noble in reason! how
infinite in faculties!"

Shakespeare: Hamlet II,2

The human brain has approximately 100 billion neurons, but it is not this capacity that gives us our noble reason. Our infinite faculties derive from the way the brain is wired through a complex network of interconnections called synapses. The number of ways this network can interconnect is not measured in billions; it is a number greater than all the fundamental particles - electrons, protons, neutrons, etc. - in the entire universe. Nature revealing the infinite within the finite.

Interconnectedness is the key to the way the human brain works and also the key to solving the precision-recall dilemma in information systems.

In the human mind words are not isolated islands like they are in machines. Every thought, word and image is intricately connected to other related words and concepts through many subtle relationships. If we want machines to be able to understand our requests for information, and to respond with comprehensive and relevant results, then we need to give them a knowledge-base that is structured the way our own brains work.

Back to Table of Contents

The Machine Correlative

The word 'syndetic' means to bind together or to connect; in information science it is applied to the cross-referencing of vocabulary terms to represent a variety of conceptual links. Standard methodologies have evolved for building relationships between concepts in information systems using controlled vocabularies. Broadly speaking these include hierarchical, equivalency and associative relationships.

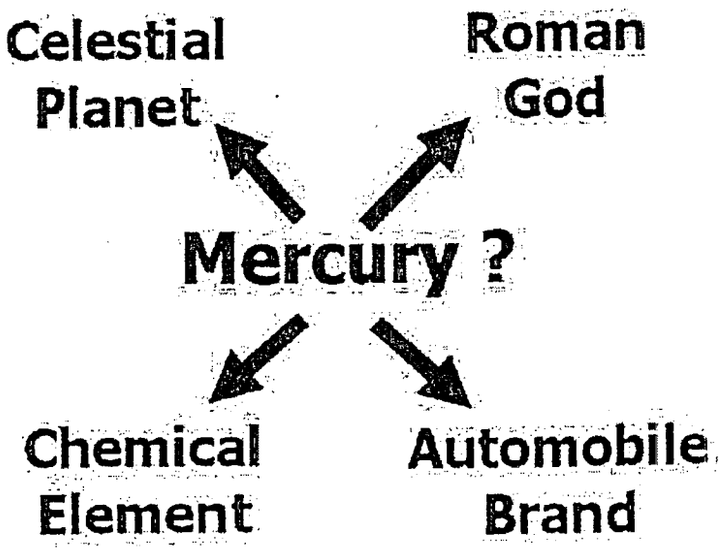
Ultimately, the goal of a controlled vocabulary is to represent each discrete real-world object or unique abstract concept by one and only one unambiguous indexing term, and then to cross-reference these indexing terms to represent the rich interconnections inherent in their conceptual or real-world correlative. Homonyms (words with more than one meaning) need to be disambiguated, and synonyms (the same meaning represented by more than one word) need to be mapped to one hub term known as the preferred term. Concepts are arranged in one or more hierarchies to represent the various categorical organizations by which the domain may be viewed. Concepts are also linked together by free association with other related concepts to further enrich the syndetic network. These types of interconnections are discussed in more detail in the following three sections.

Back to Table of Contents

The Elasticity of Language

Dr. Peter Roget, author of the first thesaurus of the English language, described a phenomenon he called 'the elasticity of language'. Put simply: one word can refer to several different things (homonyms), and one thing can be referred to by several different words (synonyms) .

For example, the word 'mercury' is a homonym. It could be used to refer to the Roman god of communication, a planet in our solar system, a chemical element, or the trade name of a line of automobiles.

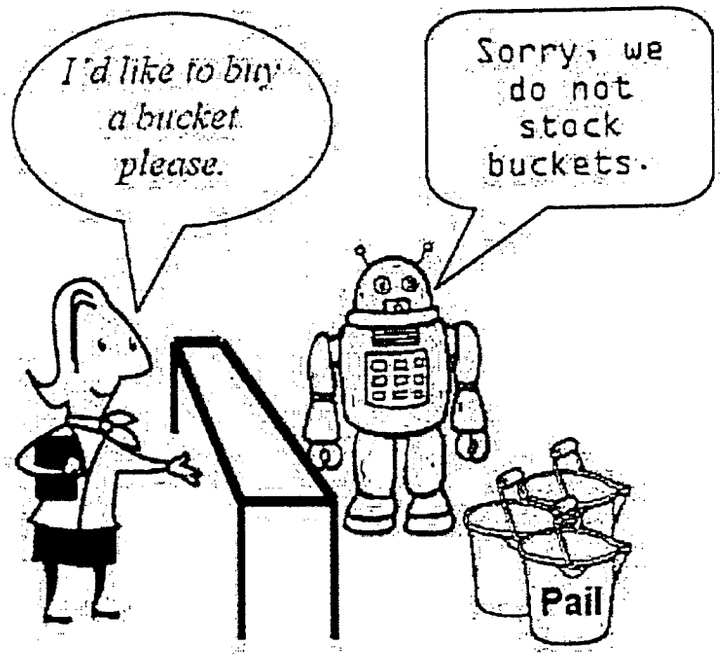


Something you might use to carry water in could variously be referred to as a 'bucket' or a 'pail', these are examples of synonyms.

If these variances seem innocuous it is because our brains instantly and subconsciously translate synonyms on-the-fly and differentiate homonyms by an internal knowledge-base of subtle contextual

associations. But the variances are far from innocuous when it comes to information retrieval systems.

Imagine going into a hardware store run by a computer and asking to buy a bucket...



The robot-shopkeeper may adamantly respond that he does not sell buckets, despite the fact that there are a dozen pails piled up in the back room. This is the most insidious consequence of the elasticity of language in text-based systems: relevant information often remains hidden. In a controlled vocabulary, equivalency relationships are used to map synonyms together and one of the synonyms is picked to be the preferred term for indexing purposes.

Less insidious but just as frustrating is the homonym problem where, for example, an online search for astrological articles on Mercury may result in the user wading through thousands (or on the web, millions) of irrelevant articles on gods, cars and chemicals. Refining the search by building a Boolean construct such as 'mercury' and 'planet' will improve the precision of the search, but comprehensiveness will necessarily suffer – some relevant results will now be excluded because they do not contain the word 'planet'. In a controlled vocabulary, homonyms are disambiguated by the use of parenthetical modifiers, e.g., 'Mercury (Planet)'; in a taxonomic structure disambiguation might be achieved by the context of the term's placement in a particular hierarchy.

Equivalency relationships may be used for other kinds of links besides connecting synonyms. They can be used for mapping spelling variants, regional and multilingual equivalents, organization-specific preferred terminology, acronyms and abbreviations.

Back to Table of Contents

Hierarchical Perspectives

Hierarchical relationships are used to classify concepts; they provide a top-down organization of ideas that present an ontology (categorical view of reality) for any given domain of knowledge.

They are often presented as taxonomies, ontologies or classifications. In a hierarchy, tree-structures are presented such that each lower level branch has some differentiated attributes while possessing the qualities of its parent branch. Differentiation can be of three types: generic, whole-part or instance.

GENERIC RELATIONSHIPS indicate that each narrower term is a kind of its parent term, e.g., cacti are a kind of succulent plant.

Plants
  Succulent plants
    Cacti

WHOLE-PART RELATIONSHIPS indicate that each narrower term is a part of (component of) its parent term, e.g., spark plugs are a part of the ignition system of an automobile.

Automotive electrical system
  Ignition system
    Spark plugs

INSTANCE RELATIONSHIPS indicate that the narrower term is an instance of its parent term. The parent term is usually a common noun category whereas the narrower term may be a proper noun, e.g., Beringer Founders' Estate is an instance of the Merlot subcategory of red wines.

Red wines
  Merlot
    Beringer Founders' Estate

Often terms may logically belong in more than one hierarchy, e.g. musical instruments might branch down paths for stringed instruments and percussion instruments; since pianos are a kind of both stringed and percussion instruments, the term pianos would therefore require placement in both structures. This is called polyhierarchy.

One of the problems of hierarchies is that there can be as many different ways of ordering a hierarchy as there are users of the information. The structure depends on the perspective of each viewer.

For example, when building a medical ontology one could identify several major views, each of which lends itself to hierarchical organization:

Diseases
abdominal diseases
cardiovascular diseases
endocrine diseases
...

Symptoms
fever
nausea
rash
...

Therapies
acupuncture
detoxification
radiotherapy
...

It would not be possible to build a single hierarchy that could structure all the permutations of these concepts in a way that would satisfy the variety of user-perspectives. In such instances it may be desirable to create a set of independent hierarchies representing each major facet of the domain of knowledge.

Facets present new challenges in user interface design as multiple hierarchies need to be navigated and dynamically recombined. One possible solution may be described using a metaphor of parallel universes and wormholes. In this model a graphical process allows the user to jump through a succession of separate ontologies or facets with a single query. The wormhole method is unique in that it presents the results as if they are a continuous hierarchical pathway, even though the actual pathway does not exist in the knowledge-base; it is dynamically constructed by the user. In the following two examples the > sign indicate where the user has wormholed from one facet to another; in both cases the results display like a continuous hierarchy.

E.g.1: Type of wine > body > wine region

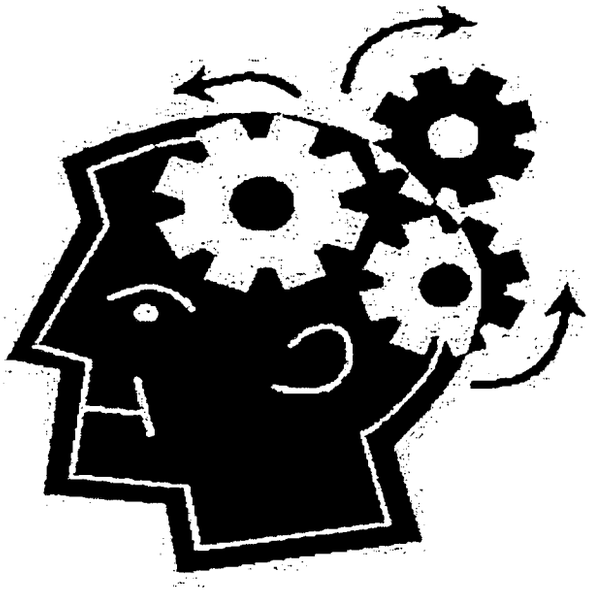
Wines
Red Wines
Merlot
> Full Bodied
> Napa Region
Beringer

E.g.2: Wine region > body > type of wine

US Wineries
Napa Region
Beringer
> Full Bodied
> Red Wines
Merlot

Back to Table of Contents

Associative Thinking



Beyond hierarchical and equivalency relationships lies the even richer and more subtle world of associative relationships. These relate concepts and objects together based upon a great variety of contextual associations. For example, temperature is related to thermometers, harvesting is related to crops, and death is related to bereavement and may even be related to its antonym life.

The links in all of the above examples represent very important associations, but none of them could be expressed through the hierarchical structures of a taxonomy nor through equivalency relationships. These associative relationships are the hardest to define, but they provide the richest and most subtle connections between concepts bringing machines closer to the kind of knowledge-base that human beings take for granted.

Back to Table of Contents

Electronic Thesauri

An electronic thesaurus represents the ultimate machine-intelligent knowledge-base for storing the myriad forms of interrelationships that can exist between concepts.

Unlike taxonomies, which only store hierarchical associations, an electronic thesaurus can combine the categorical order found in taxonomies, equivalency and mapping relationships as well as more intuitive associative links. National and international standards exist to guide the construction of such thesauri, e.g., ANSI/NISO Z39.19 and ISO 2788 and 5984.

Back to Table of Contents

Named-Entity Control

The discussion so far has concentrated on conceptual vocabularies, but controlled vocabulary techniques may also be usefully applied to other types of vocabulary, such as personal names, organizational names, geographic entities and other proper-noun lists.

Named-entity authority control is a branch of information science that utilizes the principles of conceptual vocabulary control. The goals of named entity authority control are:

1. to disambiguate different people, organizations or other entities with similar or identical names;
2. to bind together all the variant forms of names that they use;
3. to organize entities hierarchically where appropriate;
4. to display other types of association between entities.

With such a knowledge-base, machines can pull together a more complete picture about people, organizations and other entities regardless of the many variant forms of names that may be used within different information sources, and regardless of the particular form known to the inquirer. Special methods need to be employed to allow specific variant forms to be used for particular associations (e.g., Samuel Clemens was the author of certain books published under his legal name and other books under his pen name Mark Twain). Despite allowing these variant-differentiated associations, it is required to be able to cluster all associations together under a common hub for the entity.

Back to Table of Contents

Personal Names

Pseudonyms, aliases, nicknames, married / maiden names, AKAs, working names, short names, long names, code names, legalized names, pen names (noms des plumes), stage-names and sobriquets all serve to confuse and confound the proper identification of people or organizational entities in information systems. All of the above may be examples of proper and legitimate variant forms of names. The problem can be even further compounded when individuals or organizations willfully attempt to conceal their identity.

Formal and informal relationships and associations between individuals, such as marriage, parentage, friendships and business relationships, can also be expressed using customized associative relationships.

Disambiguation of individuals with similar or the same names may need to invoke rules-based validations that reference controlled metadata elements such as birthplace, residences, birth and death dates, state and other official identification numbers, and many other data.

Back to Table of Contents

Organizational Names

Organizational names apply to formal entities like corporations, academic institutions, non-profit organizations and local and national government organizations and their executive branches. They may also embrace less formal groupings of people such as performance art groups and literary circles. Covert organizations may also be covered such as criminal syndicates, political groups and terrorist cells.

Like personal names, organizational names also suffer from the confusion of aliases, pseudonyms and a plethora of other variant forms. Organizations frequently change their name through the process of mergers, acquisitions and partnerships. Many organizations can be arranged with some hierarchical structure such as the corporate division / subdivision structure.

Important associations between organizations can be identified such as Company A merged with Company B, Company C customer of Company D, etc.

Personal named entities may also be associated with organizational named entities to show the membership of specific individuals in particular organizations.

Back to Table of Contents

Geographic Names

Geographic names apply to countries, states, regions, cities, subdivisions, territories, seas, rivers, mountains, deserts, forests, jungles, plains, and many other geographic and topographic entities. These entities also have variant forms of names, identical names are duplicated in different places and the names evolve through history with shifts in territorial boundaries.

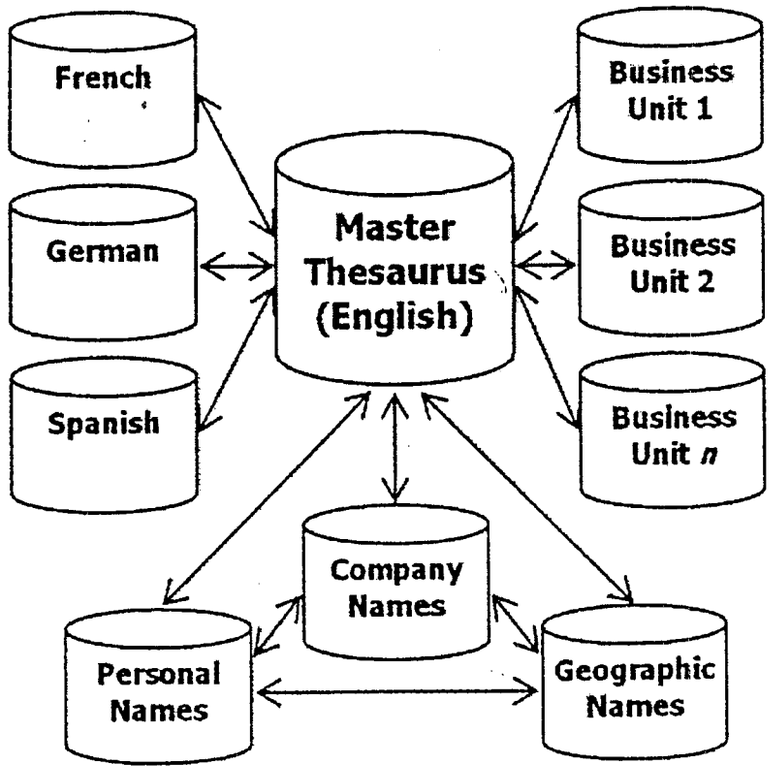
Most geographic entities lend themselves to hierarchical structuring as well as other associative links.

Back to Table of Contents

Enterprise Knowledge-Bases

For most large corporations and government organizations the goal is to build a master controlled vocabulary that unifies all the data sources in the enterprise. The enterprise might have some business units that use specialized, private, legacy or third-party vocabularies. By mapping these satellite vocabularies to the hub of a master vocabulary the organization can benefit from cross-searchability throughout the entire enterprise. This configuration may be called a metathesaurus.

Following is a diagram of a model enterprise-wide knowledge-base. Multilingual meta-thesauri, private and third-party vocabularies and named-entity vocabularies are all mapped through a single syndetic network. Note: all conceptual vocabulary, including multilingual versions, map to a common hub (the master thesaurus); the master thesaurus and the named entity files are all mapped to each other.



Back to Table of Contents

Six Degrees of Separation

The phrase six degrees of separation was coined by John Guare in his 1991 stage play of the same name. The concept is that any one can make a connection to any other person on earth through a chain of six intermediary people; each link in the chain must be between people personally acquainted or familiar with each other. At first the proposition sounds unbelievable, but the play was inspired by empirical research undertaken by Harvard professor Stanley Milgram in 1967. Milgram's research was based on connecting two people randomly picked from anywhere in the USA; his study resulted in an average of 5.5 intermediary people.

What we are observing in the six degrees of separation phenomenon is the close proximity between nodes in complex networks, and how fast and easy it is to get from one place to another in remarkably few leaps (phone calls, mouse clicks, etc.).

Back to Table of Contents

Conclusions

Text-based searching necessarily fails to yield accurate and comprehensive results due to the inherent elasticity of language. The human brain manages to overcome this problem because humans have a sophisticated knowledge-base in which no word is an island. In order for machines to be able to achieve the same level of precision and recall when searching for information, they need controlled vocabularies with rich syndetic networks.

Scale is important - small and simple taxonomies are not sufficient for accessing the rapidly growing information repositories managed by large corporations and government organizations.

Syndetic networks, however, can disambiguate similar concepts and names, pull together variant forms, capture equivalencies between multilingual vocabularies, arrange concepts and named entities into multiple hierarchical groupings, and capture subtle associative relationships between concepts and named entities.

The richer the network's cross-referencing within and between these vocabularies, the more successful will be the syndetic network at answering user queries accurately, comprehensively and quickly.

"All around us are facts that are related to one another. Of course, they can be regarded as separate entities and learned that way. But what a difference it makes when we see them as a pattern! . . . They begin to make sense. The world becomes a more comprehensible place."

MURRAY GELL-MANN

References and Resources

Click on References and Resources for external links to useful resources and references on

controlled vocabularies.

Copyright © 1995 - 2004, by Synapse, the Knowledge Link Corporation, on the web at www.synaptica.com and www.taxonomywarehouse.com. Synaptica ® is a Registered Trademark of Synapse, the Knowledge Link Corporation. Page created 10/9/2004 - 6:41:27 AM US Mountain Time.

Relevance: Language, Semantics, Philosophy

JOHN M. BUDD

ABSTRACT

THE LITERATURE WITHIN LIBRARY AND INFORMATION SCIENCE (LIS) on relevance comes primarily from the subfields of information retrieval and information systems design. This literature has developed over time from an orthodoxy that has focused on relevance as an objective measure to a comprehension of the dynamic nature of relevance judgment. Other literatures, such as those of the philosophy of language and semantics, also have offered cogent thought that could and should be incorporated into LIS. This thought has broadened discussion to the context in which relevance is assessed, the speech acts that are evaluated, and the dialogic element of human communication.

An individual may use any number of ways to begin an examination of relevance. One beginning is provided by Fred Dretske in his *Knowledge and the Flow of Information*. Dretske acknowledges the usefulness of quantitative theory applied to information but asserts that such theory is limited in that it cannot elucidate the nature of meaning or tell us about the meaning of a particular statement. He says that "if we consult a dictionary, we find information described most frequently in terms of 'intelligence,' 'news,' 'instruction,' and 'knowledge.' These terms are suggestive. They have a common nucleus. They all point in the same direction—the direction of truth. Information is what is capable of yielding knowledge, and since knowledge requires truth, information requires it also" (Dretske, 1981, p. 45). Dretske's notion may make us wonder about the connection between information and knowledge.

The role of relevance as it relates to knowledge will recur in this pa-

John M. Budd, School of Information Science and Learning Technologies, University of Missouri-Columbia, 303 Townsend Hall, Columbia, MO 65211
LIBRARY TRENDS, Vol. 52, No. 3, Winter 2004, pp. 447–462
© 2004 The Board of Trustees, University of Illinois

per, but some background should guide this exploration of the philosophy of relevance. Fortunately for us, Stefano Mizzaro (1997) undertook an exhaustive review of the information science (IS) literature relating to relevance. His work renders a descriptive literature review here unnecessary; readers should consult his article and its extensive bibliography. In setting the tone for his review, Mizzaro acknowledged that IS approaches to relevance tend to cluster around two groups—one centering on the object, or bit of information, and the other centering on the human element. The first group (the object cluster) includes three entities according to Mizzaro: a document, a surrogate (or representation of a document), and information (or what the reader apprehends from a document). The second group includes a problem faced by the information-seeker, an information need (defined as a mental representation of the problem), a request (a natural language expression of a need), and a query (or system language expression of a need) (p. 811). He then posited, "Now, a relevance (*sic*) can be seen as a relation between two entities, one from each group: The relevance of a surrogate to a query, or the relevance of a document to a request, or the relevance of the information received by the user to the information need, and so on" (p. 811).

We can take for granted, for the purposes of this examination, that Mizzaro's observations regarding relevance provide a reasonably accurate and accepted summary of IS inquiry and system development work. They do, however, raise some questions for more broadly defined philosophical treatment of relevance. Inherent in his set of clusters (particularly for the first, object-based, group) is the assumption that relevance applies primarily to verbal communication. Further, the assumption is that this communication is formal and structured, that is, it can be shaped and presented in the form of a document. Granted, "document" can be an encompassing idea, but a formal structure inheres in it, even if it is not intended to be limited to a physical artifact that would satisfy a popular notion of what a document is. An assumption that is evident with regard to the second cluster is that humans initiate a process whereby information is sought and located. Further, the two clusters combined (as Mizzaro did to define relevance) suggest a structured human action that entails the making of specific kinds of judgments about objects by people. This model is not particularly problematic (on the face of it) as far as it goes, but it does not go far enough. There are information-related human actions that do not fit so neatly into the clusters, much less into the individual elements of the clusters.

This very brief bit of background serves to demonstrate some of the complications one faces when examining relevance. There is no way that a paper of this length can possibly address all definitions, uses, and implications of the word relevance; what is presented here is a selection of some ways of thinking about relevance. One point needs to be made immediately: whether stated or not, relevance judgments are fundamentally construed

in epistemological terms, that is, the aim of a relevance judgment is, in the end, to foster or justify knowledge. First, a few approaches from library and information science (LIS) will be discussed. Following that, some potentially fruitful philosophical ideas will be dealt with. The desired result is (i) a realization of the complexity of any notion of relevance; (ii) an understanding that different starting points for an examination of relevance may well lead to different conclusions; and (iii) while philosophers do not appear to address relevance directly, many do give serious attention to matters that impinge on our understanding of relevance.

LIS AND RELEVANCE

While Mizzaro provided a succinct summary of IS writings and thought on relevance, a few specific works from the past have been especially influential or have provided examples of particular schools of thought regarding relevance. The first, and what for many readers may be the most striking, aspect of the LIS literature is that the preponderance of writing on relevance comes from IS (including the subfields of information systems design and information retrieval). Very little is present in the literature about libraries that directly addresses relevance (apart from systems-related concerns). In works on the reference function in libraries, for instance, a tacit assumption is that the process is intended to help library users find materials and information that they will find useful, but little reference to writings on relevance itself exists. It could (and probably should) be argued that greater attention to the complexity of relevance and how individuals may make relevance judgments is vitally important to the essence of the library's reason for being. Formal examination of relevance, both as an idea and as a phenomenon, tends to be found in such journals as the *Journal of the American Society for Information Science and Technology* and *Information Processing and Management.* Whether practicing librarians read these journals to learn more about things like relevance is an open question.

Given that the literature of IS tends to be the home of discourse on relevance, it is essential that the substance of that discourse be considered here. It should be noted with some emphasis that what follows is a focus on some particular, but representative, ideas about relevance; what is covered here is by no means exhaustive. While Mizzaro did review a large amount of literature, a few rather consistent themes recur. Moreover, the themes seem to have a kind of temporal aspect, that is, at one point in time a sort of orthodoxy prevailed, and over time alternative conceptions of relevance have been articulated. The orthodoxy tends to fit into Mizzaro's first cluster that centers on the document (and the information system). About four decades ago William Goffman's approach expressed a particular, and apparently widely held, view of relevance assessment. In his article he (Goffman, 1964) wrote that "relevance can be defined as a measure of information conveyed by a document relative to a query" (p. 201). There

are some conceptual and practical difficulties with such a view. For one thing, nowhere in the article is "information" defined; for another, a deterministic theorem and proof are presented as rationale for the measurement of relevance (recall Dretske's skepticism regarding the sufficiency of quantitative theory). Perhaps more fundamentally, relevance is explicitly referred to *as* a measure. The assumption underlying the claim is that relevance is a physical, or at least a tangible, thing that can be assigned a quantitative value, and that value can be related to the value of other variables, such as information.

A similar way of defining relevance was expressed a bit later by W. S. Cooper (1971). He said, "A stored sentence is *logically relevant* to (a representation of) an information need if and only if it is a member of some minimal premiss (*sic*) set of stored sentences for some component statement of that need" (p. 24). Inherent in this definition is a very specific equation of relevance with representation; it embodies what Alvin Goldman (1993) called a representational heuristic. This is "the tendency to judge the probability that an object x belongs in category C by the degree to which x is representative of, or similar to, typical members of category C" (pp. 26–27). Such a heuristic may be entirely effective for certain cognitive processes and informational needs, such as the search for documents or statements that have direct bearing on a claim. Suppose a student or a scholar is examining a claim that an economic policy decision was made in a particular U.S. presidential administration for the primary purpose of helping win votes in an upcoming election. That person would probably seek documents that address the claim rather directly, that is, that discuss the political implications for the specific policy decision. In such an instance, the heuristic described by Goldman holds. Suppose, however, we consider the description of a method of inquiry by the historian William H. McNeill: "I get curious about a problem and start reading up on it. What I read causes me to redefine the problem. Redefining the problem causes me to shift the direction of what I'm reading. That in turn further reshapes the problem, which further redirects the reading. I go back and forth like this until it feels right, then I write it up and ship it off to the publisher" (Gaddis, 2002, p. 48). McNeill's experience describes a different kind of cognitive process; in his account what Goldman refers to as categorization is dynamic, not fixed. Because the category is dynamic, representation is likewise dynamic and shifting. No linear algorithm is sufficient to account for McNeill's process.

Following Cooper were the beginnings of alternative conceptions of relevance. The first expressions of alternative discourse were still substantively grounded in the orthodoxy, but some discomfort with that orthodoxy seems apparent. Building upon Cooper's work, Patrick Wilson (1973) added a refinement, which he called situational relevance. The situation, or context, within which an information-seeker assesses the relevance of a document or an utterance implies a logical functioning that differs from

the one Cooper suggested. In addition to what may begin as a deductive process, iterative steps of reading, inference, and continued searching necessitate, according to Wilson, probabilistic evaluation of each item plus induction stemming from the reading of those retrieved items. On the face of it, Wilson's addition of these stages appears to be only a small twist to the structure that Cooper, Goffman, and others posited. In actuality, the logical complication Wilson interjected begins to resemble more closely the complex process summarized by McNeill. The melding of deduction and induction (and McNeill's anecdote clearly supports this melding) stresses that the ultimate judgment of relevance in all but the most simple instances is a nonlinear and rich process.

Abraham Bookstein (1979), following Cooper and Wilson, added some potentially complicating ideas to relevance assessment. He admitted that "an information retrieval system cannot predict with certainty a patron's reaction to a document, and this, we believe, is the source of many of the uses found in the literature for the term 'relevance.' [At this point Bookstein seems to be migrating from that cluster of the literature centering on the document to the cluster dealing with the user, but the continuation of his thought suggests otherwise.] Rather, the system transforms both the document and the request into forms it can manipulate, and on the basis of these, it *assesses* the relevance of the document to the user" (p. 269). He further defined relevance in terms of a user's satisfaction with the output of a system. This idea begs some questions, including what constitutes satisfaction (is it in fact the retrieval of topical, or more narrowly, linguistically related, documents) and how does an information retrieval system accomplish the goal of satisfaction. These are not trivial questions; if a user's query is reduced to linguistically morphological form (and it must be added that sophisticated systems approach more semantically related goals), then assessment is a simplified, but far less meaningful, task. Since satisfaction is not uncommon in library-related discourse, the questions (and the associated problems) also apply.

Some other IS writings introduce some evaluative mechanisms that can reinforce the objective idea of information. These measures recur in several sources, but two examples of their statement will suffice for illustrative purposes. In her book on information retrieval, Miranda Pao (1989) summarized two measures that are intended to help assess relative effectiveness and utility of searching for and retrieving information. These two measures are (i) recall (the number of relevant items retrieved divided by the number of relevant items in the database) and (ii) precision (the number of relevant documents retrieved divided by the total number of retrieved items) (p. 59). These measures are problematic inasmuch as they either leave "relevance" undefined or assume that items and documents can objectively be categorized as relevant or not relevant (with little or no middle ground). A more fundamental underlying assumption imbedded in such instrumental mea-

sures is that relevance, once a search is expressed, is a property that inheres in the document retrieved. There may be an admission that some information-seeker is making a relevance judgment about the members of a set of documents, but there remains the operational procedure of treating relevance as something that is *of* the document. The measures also are referred to by Michael Buckland (1991, p. 101). Buckland, however, recognized the semantic challenges of claiming that a relevance judgment is an objective one. While admitting that ascribing relevance to a document, as part of the execution of the recall and precision measures, is a matter of treating information as a thing, he readily admitted that information also can be perceived as knowledge and as a process. Both of these conceptions transcend the limited view of relevance as objective.

The shift from orthodox to alternative ideas concerning relevance is well-illustrated in an article by Linda Schamber, Michael Eisenberg, and Michael Nilan (1990). They reviewed traditional and nontraditional conceptions of relevance judgments and demonstrate a shift from static to dynamic factors. In summary they concluded that

1. Relevance is a multidimensional cognitive concept whose meaning is largely dependent on users' perceptions of information and their own information-need situations.
2. Relevance is a dynamic concept that depends on users' judgments of the quality of relationships between information and information-need at a certain point in time.
3. Relevance is a complex but systematic and measurable concept if approached conceptually and operationally from a user's perspective. (p. 774)

The third point relies on confidence in formal relevance judgment as a rational process. While some might challenge such a view, its debate is beyond the scope of our concern here. One thing that is of some importance here is the nature of knowledge itself. A reductionist view of knowledge would hold that knowledge is subject to either an internalist (all knowledge resides inside the individual knower's mind) or an externalist (all knowledge is grounded outside the knower's mind) stance. It seems clear that the orthodox view of relevance is externalist; assessments of the relevance of specific documents are possible primarily because of properties of the documents themselves.

The third point articulated by Schamber et al. presents a difficult, but essential, characteristic of relevance. If we were to take a purely externalist stance regarding knowledge, and if we make relevance analogous to knowledge, then relevance would necessarily be external to the information-seeker, and researchers could conceivably evaluate the relevance of documents. If, on the other hand, we were to take an internalist stance regarding justification (the grounds for generating belief about a claim or statement), and

if relevance were analogous to justification, then relevance would necessarily be within the information-seeker. Rober Audi (1998) in his book on epistemology convincingly argued that knowledge and justification are, to an important extent, separate and that an externalist view of knowledge combined with an internalist view of justification is legitimate, and that, further, justification is important to knowledge (pp. 237–238). Taking a cue from Audi, it may be that we need to look at relevance as dually external and internal to the information-seeker. Some object is read and evaluated, but the evaluation (justification) is an outcome of rational, internal processes. In other words, some meaning inheres in a document (external) and the meaning is contextualized and assessed by a user (internal).

Some recent work in IS incorporates the complex epistemological notion of combined internalist and externalist factors. Stephen Harter 1992) recognized the methodological challenges of discarding ideas of fixed relevance but asserts that matters of topicality are less important than assessment of cognitive change in the information-seeker. Harter drew, in part, from work on the cognitive elements of relevance as set forth by Dan Sperber and Deirdre Wilson (1986), about which more will be said later. In his introduction to a special issue of the *Journal of the American Society for Information Science,* Thomas Froehlich (1994) offered a particular means of broadening the discourse on the matter, tacitly incorporating the dualist principle suggested by Audi. Froehlich wrote, "Hermeneutics can provide a productive framework for modeling systems and user criteria" (p. 130). He described some of the factors that impinge on a hermeneutics of relevance, including the realization that users interpret their needs to and for themselves, understanding how documents become part of an information collection such as a library, and determining how a user's query in interpreted by and through an information system (p. 130). The addition of the interpretive element was not entirely new with Froehlich; hints of it were offered by David Ellis (1984) who noted the need to take into account the perception of the user and the impact of that perception on relevance judgments. With the introduction of hermeneutics, Froehlich made apparent how the interpretive process can itself be examined, without ignoring the intentional nature of system creation and development.

One of the most sophisticated assessments of relevance as a concept was that of Birger Hjørland. In one of his works, Hjørland (1997) reminded us that awareness of relevance coincides with an explicit awareness of information need; the criteria to evaluate the relevance of a document are ineluctably attached to the need itself (p. 172). In this the simultaneous internal and external elements are clear. Hjørland's thought (2002) is most helpful when he demonstrates that the conception of relevance is a manifestation—explicit or not—of a particular school of thought. The criteria for relevance are derived from the epistemological framework within which the researcher works. For example, if the researcher is an empiricist, rele-

vance is related to observations, sense data, and not from outside authority or testimony (p. 117). It is essential to remember that the decisions regarding criteria for evaluation of relevance are grounded in a theoretical stance, whether the theory is articulated or not. A criticism that emanates from many quarters (many disciplines) is that the theory is, in general, far too frequently unstated, even unrecognized. The sub-rosa aspect of theory in LIS leads to problems of conceptualization, definition, assumption, evaluation, and, ultimately, understanding.

The View from Without

It is impossible to consider relevance without delving fairly deeply into Sperber and Wilson's (1986) book. Theirs has been the most substantive and complete consideration of the cognitive elements of relevance as they relate to communication to date. LIS should pay attention to a few key aspects of their position. One aspect (although not the most major one) is that, in the communication process, interpreters are concerned with utterances (complete, or nearly complete, statements, arguments, propositions, etc.), rather than sentences. This seems minor, but it emphasizes that, as people hear or read, they usually are not simply extracting individual and small parts of a discourse; they are evaluating larger communicative acts. In customary library and information-retrieval settings, individuals tend to be making judgments about documents or substantial portions of documents.

Of greater interest here is an aspect that is more central to Sperber and Wilson's concept of relevance—context. They defined context as "a psychological construct, a subset of the hearer's assumptions about the world. It is these assumptions, of course, rather than the actual state of the world, that affect the interpretation of an utterance" (1986, p. 15). Anyone considering relevance seriously must heed the first part of this definition (the hearer's or reader's assumptions about the world). The second part, if not examined further, could lead us down an unproductive path, though. It is also essential that Sperber and Wilson wrote that context is determined by *more than* the actual state of the world. If one were to interpret their definition as meaning that context is independent of the world, then this would represent an antirealist view of the world. A realist, even a weak realist, view would necessitate that we accept that the state of the world, at the very least, influences our assumptions. To affirm that their position is not a completely antirealist stance, they claim that mutual knowledge is vital to any relevance assessment and, so, to communication. For mutual knowledge to be possible, there must be shared assumptions, which are most likely to come from the common influence of the state of the world. The idea of mutual knowledge applies in many information-related instances; the creator of a document/utterance frequently has an audience (though sometimes an ideal audience) in mind. The utterance, then, may be made with the knowledge base of the audience taken into account. To the extent that there genuine-

ly is mutual knowledge, there is the likelihood that members of that audience may find relevance in the utterance.

What is most fruitful in the development of an idea of relevance by Sperber and Wilson is their outline of the cognitive processes involved in the assessment of relevance. This assessment depends on context, as defined above. Complicating the cognitive evaluation of relevance is the realization that any hearer/reader constantly resides within the framework of multiple contexts. This becomes obvious if we consider the actions of a student or scholar living in a rich environment. At any given point in time, the individual is likely to be working on more than one project. Even when the individual is consciously and intentionally trying to focus on one context (say in conducting a search of a literature with the possible connection with one project in mind), the other contexts do not simply disappear. To the extent that the other contexts become conscious, the literature being assessed may be assessed within more than one context. If the multiple contexts are related by content, then the assessment may be further complicated. The individual reviewing literature may read a given abstract to determine if the related paper is relevant to the project. Something in the abstract, however, is semantically connected to a second project, which shifts the individual's contextual state; the individual begins to read the abstract with the second project in mind. It may be that, if there is a content relation between the two projects, the assessment of the abstract within the context of the second project affects the context of the first project. In other words, as some writers within LIS have noted, relevance judgment is unavoidably dynamic.

Sperber and Wilson explicitly observed that any individual may be assessing intuitions relating to relevance in multiple, and shifting, contexts, and it is difficult, if not impossible, for a third party to be certain within which context assessment resides at any given moment. They wrote, "As a discourse proceeds, the hearer retrieves or constructs and then processes a number of assumptions. These form a gradually changing background against which new information is processed" (1986, p. 118). Sperber and Wilson offered another observation that is essential to a full consideration of relevance—relevance judgment tends not to be a binary decision; rather, it is an assessment made on a continuum. This means that the decision is a relative one; utterance A may be deemed relevant but less relevant than utterance B. One additional point by Sperber and Wilson (1986) must be considered: "We have suggested that the context used to process new assumptions is, essentially, a subset of the individual's old assumptions, with which the new assumptions combine to yield a variety of contextual effects" (p. 132). Relevance judgment is dynamic because context is dynamic because assumptions are dynamic. New assumptions may be formed both on the basis of old assumptions and leading to a transformation of old assumptions. As the transformation takes place, the judgment of the relevance of a particular utterance is subject to change.

The points made by Sperber and Wilson are useful and informative as we in LIS consider relevance more fully. Their thought, of course, did not spring fully formed without being influenced (contextually and via assumptions) by other writings and utterances. Two of these influential utterances (one tacit and one explicit; that is, one not referred to in their book and the other one included in their bibliography) deserve brief treatment here. The tacit influence is Ludwig Wittgenstein's *Philosophical Investigations* (1958). In part, Sperber and Wilson appeared to be reacting against some points by Wittgenstein. For example, Wittgenstein (1958) said, "A proposition, and hence in another sense a thought, can be the 'expression' of belief, hope, expectation, etc. But believing is not thinking. . . . The concepts of believing, expecting, hoping are less distantly related to one another than they are to the concept of thinking" (p. 154). His claim raises the question of whether relevance judgment falls in the realm of belief or of thought. The answer has decided implications for any inquiry into relevance. It seems apparent that Sperber and Wilson did not follow Wittgenstein on this matter; they turned, in part, to some works by Jerry Fodor to support an idea of greater coherence between such things as belief and thought. Fodor (1975) maintained that meaningful use of language necessitates at least some correspondence between an expressible belief and the actual expression (in language) of the belief (p. 72). Wittgenstein, though, posited a notion that requires addressing, even if there is no specific reference to his work. Curiously, this example could be taken to be indicative of the complexity of relevance. While Wittgenstein was not cited in Sperber and Wilson, his work is relevant, even if negatively in this singular instance, to the program Sperber and Wilson set forth.

The explicit influential utterance referred to above is that of John Searle (1983). Two points made by Searle suffice to illustrate both the indication of influence on Sperber and Wilson and some essential considerations relating to relevance. First, Searle addressed the matter of communication from the standpoint of the speaker. If an utterance by a speaker is to be meaningful, then the speaker must have had a set of intentions directed at a set of hearers, at an audience. The set of intentions entails having an effect on the audience (Searle, 1983, p. 161). The importance of this point is to remind us that, as an individual judges the relevance of a document (utterance), the judgment is influenced by the document itself, especially whether, at that moment, the individual falls into the category of intended audience of the creator of the document. If the answer is yes, then there is a higher probability (though almost impossible to calculate) that there is a contextual connection between the individual and the document; the individual's assumptions are related to the content of the document. Stated in Searle's terms, a connection exists between the set of intentions held by the individual seeking information and that set held by the speaker/author at the time of the uttering.

Another point by Searle addressed the question regarding where relevance resides. Sperber and Wilson emphasized their perceived importance of an individual's assumptions about the world (as opposed to the state of the world). Searle (1983) reviewed the distinction (in the mind, or *de dicto* beliefs, and beliefs that are about real objects, or *de re* beliefs). While the evaluation of beliefs may be difficult, he concludes that there can be both *de dicto* and *de re* beliefs (pp. 208–210). If an individual is reviewing a retrieved set of documents, that individual may read one that states putative facts about, say, cognitive processes employed in information retrieval. The individual also may be reading another that argues for a particular kind of behavior to be employed in reference transactions, based on claims of the emotional state of the inquirer. An understanding of the assumptions underlying assessment of these two documents requires attention to the nature of content of the documents. Assessment of the content is itself subject to kinds of epistemic justification. The one document may be deemed justified on the grounds of physical evidence of cognitive processes; the other may be deemed justified on the grounds of effective argumentation. Clearly in Searle, but also apparent in Sperber and Wilson, both kinds of beliefs are integral to relevance judgment.

Searle's insights are certainly valuable for any consideration of relevance, but it should be noted that his position represents one of at least a few possible stances regarding meaning. For any document or utterance to be judged relevant, it stands to reason, that document or utterance should be deemed meaningful by the reader or hearer. Speech acts, which form the core of Searle's concern, can be assessed as potentially meaningful. These speech acts, however, tend to be assessed on the grounds of the matter or content they communicate (a second conception of meaning). Paul Grice (1989) demonstrated that the kinds of communication that generally are considered when attention is turned to relevance (that is, formal assertions based on evidence or logic) are accompanied by particular communicative intentions. The intention is usually to communicate the substance of the idea or thought that the speaker/writer has to an audience. Grice's position is dependent on an even more fundamental notion of meaning (a third conception) as an explanation of "what it is to think that P, what it is to believe that P, to desire that P, etc." (Harman, 1999, p. 158). This notion attempts to theorize on the nature of a language of the mind that allows for communication that can be comprehended internally by the thinker and then communicated to someone else. While this third conception is important, the first two have more direct relationship to an understanding of relevance.

Borrowing from another work from outside LIS, while systemic concerns are present within our field for the consideration of relevance (i.e., answering the question of whether the information system—database, service, etc.—meets the information need of the inquirer), it may be that "sys-

tem" is even more useful metaphorically. Patrick de Gramont (1990) took this tack when he compared the workings of language to a filing system. He wrote:

> Filing systems have two distinguishing characteristics which enable one to compare them to the way language works. First, they operate on the basis of the fact that the information to be filed has meaning before it is filed. Second, the system under which the information is filed is geared, not to the information per se, but to an ulterior purpose. For example, if I file my correspondence alphabetically, the classification I use has nothing to do with the correspondence in itself; rather it is a function of wanting to retrieve letters easily and efficiently. (p. 65)

De Gramont hastened to add that while this metaphor effectively illustrates the ways we employ language for particular purposes, purposes may become conflated and render meaning difficult. De Gramont (1990) said that the meaning of a file (that is, the categorical meaning it has so that items can be attached to it) is not the same as the meaning of the content within the file; the categorization, being an organizing technique, is ineffective as a meaningful indicator of specific content (p. 69). One reason for the difficulty is that the employment of language for the purpose of categorization (filing) tends to be an application of process rules, and not necessarily related to meaning.

De Gramont's metaphor maps onto our disciplinary concern regarding relevance. The system-based work on relevance that characterizes some IS literature is, if de Gramont is correct, not only limited, but also misleading. If the initial concern is determination of the relevance of documents retrieved from an information system (including a library), then the vital matter of the meaningfulness of possible connection of the intention of the searcher, the stated search, and the retrieved output is forgotten. As the process moves from stage to stage, some transformation, along the lines of the metaphor of the filing system, occurs. The searcher's intention is translated into a query, which is further transformed through the mapping of search terms onto document representations. Serious inquiry into relevance judgment should not ignore the transformative processes that take place along the way. Applying de Gramont's thought to relevance might prompt the profession to think seriously, apropos of Mizzaro, about the two clusters of work on relevance, and especially whether the separation into two clusters not only misses the point of relevance but also possibly perpetuates a mistaken notion of system and human operating independently. This is not to say that there is no awareness that the two clusters are really only two emphases within an interrelated phenomenon; there is certainly work that does integrate the two.

One additional view from the outside can help demonstrate the more particular concerns associated with relevance judgment that may be a part of a dialogic process (e.g., reference transactions). While a number of think-

ers could enlighten us, one philosopher in particular has addressed matters of communicative action—Jürgen Habermas. In one specific work Habermas (1998) has critiqued some standard theories of meaning and has found them wanting. A primary difficulty with these theories is that they tend to be constituted almost entirely in language itself and not in what Habermas referred to as the pragmatic relationship between speakers and hearers that can be both linguistic and extralinguistic (p. 280). What is missing in such a case is a lack of attention to action—theoretic contexts in which meaning may be realized. As mentioned above, Birger Hjørland in our field has attempted just such a connection between meaning and action theory. According to Habermas, traditional theory has not fully abandoned a certain articulation of semiotics wherein an object-centered idea of knowledge holds that meaning (signified) relates to a sign (signifier) in the same limited way that a potentially meaningful sign (symbol) relates to the signified object (designatum) (p. 281).

Habermas offered a replacement for the traditional theory by emphasizing the tripartite objective of "a *speaker* reaching understanding/ with *another* person/ about *something*" (p. 293). He added to the dynamic combination of *de dicto* and *de re* beliefs affecting meaning by explicitly recognizing the shared social world in which communication resides (p. 296). His addition of this component has the effect of clarifying that, rather than being a linear, objectified process, communication (and relevance judgment) is a dynamic and transformative force that can enable understanding to take place.

> The telos of reaching understanding inherent in the structures of language compels the communicative actors to alter their perspective; this shift in perspective finds expression in the necessity of going from the objectivating attitude of the success-oriented actor, who seeks to *effect* something in the world, to the performative attitude of a speaker, who seeks to *reach understanding* with a second person about something. (Habermas, 1998, p. 300)

Any dialogic effort to determine the relevance of documents or utterances depends on the kind of teleological stance of which Habermas speaks. This means that, from the professional point of view, it is not sufficient to recognize a simply stated goal of service; a more robust and rich articulation of purpose is required to meet the goal of finding relevance in a dynamic and complex environment.

Summary

In recent writings, representatives from the IS field have recognized the dynamic and complex nature of relevance and relevance judgments. Some of these writers go so far as to delve into work beyond LIS. The reality, however, is that such treatments are limited. Furthermore, consideration of relevance is to be found in the IS literature, rather than in the literature

on librarianship. As is the case with any absence, reasons are left to speculation. It may be that librarians read the IS literature and learn from it. It may be that librarians tacitly believe that practice effectively embraces relevance in the form of reference service and measures such as user satisfaction. When we traverse beyond our own literature, we can readily see a richness of thought relating to language, context, content, and other factors that are closely connected to relevance. The potential for understanding and misunderstanding may be the most fundamental of ideas that can guide the inquiry and practice in which relevance holds a central place.

When Habermas's critique of theories of meaning is considered, the challenge of understanding what relevance is and how it can be assessed becomes clear. Habermas, more so than many other writers on language and communication, comprehends the inherently dialogic nature of communication. To ground an examination of relevance even more explicitly in dialogic communication, the work of Mikhail Bakhtin should be considered. Bakhtin, more effectively than anyone else, has captured the elusive quality of dialogue. Underlying his program is the claim that all communication is dialogic or monologic. Monologic communication allows no response, no appropriation by a reader or hearer; it simply is as it is stated. Dialogic communication requires interaction between hearer and speaker, between writer and reader. This kind of communication also requires the admission that "Language is not a neutral medium that passes freely and easily into the private property of the speaker's intentions; it is populated—overpopulated—with the intentions of others. Expropriating it, forcing it to submit to one's own intentions and accents, is a difficult and complicated process" (Bakhtin, 1981, p. 294).

We may be tempted, especially given the force of a narrow empiricist tradition in our field, to presume that a relevance judgment is not subject to multiple, sometimes competing, intentions, that the information-seeker imposes her intentions on the document/utterance. The full realization of relevance entails what Michael Bernard-Donals (1994) called the epistemological foundation of Bakhtin's work—"the I–other relationship that takes place between humans through the creation of signs (and more specifically, with language).... [H]umans are radically 'other' in relation to each other, but it is this relationship that defines human understanding, and all epistemologies must come to terms with it, as Bakhtin's does" (p. 43). As is stated early in this paper, relevance is necessarily connected to knowledge. An understanding of the phenomenon of relevance also necessitates examination that transcends a linguistics analysis that ignores the dialogic nature of communication. Bakhtin (1986) summed up what ideally occurs in relevance assessment:

> The transcription of thinking in the human sciences is always the transcription of a special kind of dialogue: the complex interrelations be-

> tween the *text* (the object of study and reflection) and the created, framing *context* (questioning, refuting, and so forth) in which the scholar's cognizing and evaluating thought takes place. This is the meeting of two texts—of the ready-made and the reactive text being created—and, consequently, the meeting of two subjects and two authors. (pp. 106–107)

Any dynamic conception of relevance requires an understanding of the dialogic, and essentially phenomenological, process of communication. As Emmanuel Levinas (1969) has told us, discourse involves the production of meaning, not merely an ideal meaning, but one that is grounded in the present act of reading/hearing the words of the writer/speaker (p. 66). A judgment of relevance is likewise grounded in the present. This brief examination demonstrates that fruitful thought from philosophy, language, and semantics can help LIS delve more deeply into the study of relevance.

References

Audi, R. (1998). *Epistemology: A contemporary introduction to the theory of knowledge.* London: Routledge.

Bakhtin, M. M. (1981). *The dialogic imagination: Four essays* (C. Emerson & M. Holquist, Trans.). Austin: University of Texas Press.

Bakhtin, M. M. (1986). *Speech genres and other late essays* (V. W. McGee, Trans.). Austin: University of Texas Press.

Bernard-Donals, M. F. (1994). *Mikhail Bakhtin: Between phenomenology and marxism.* Cambridge: Cambridge University Press.

Bookstein, A. (1979). Relevance. *Journal of the American Society for Information Science, 30*(5), 269–273.

Buckland, M. (1991). *Information and information systems.* New York: Praeger.

Cooper, W. S. (1971). A definition of relevance for information retrieval. *Information Storage and Retrieval, 7*(1), 19–37.

de Gramont, P. (1990). *Language and the distortion of meaning.* New York: New York University Press.

Dretske, F. I. (1981). *Knowledge and the flow of information.* Cambridge, MA: MIT Press.

Ellis, D. (1984). Theory and explanation in information retrieval research. *Journal of Information Science, 8*(2), 25–38.

Fodor, J. A. (1975). *The language of thought.* Cambridge, MA: Harvard University Press.

Froehlich, T. J. (1994). Relevance reconsidered—Towards an agenda for the 21st century: Introduction to a special topic issue on relevance research. *Journal of the American Society for Information Science, 45*(3), 124–134.

Gaddis, J. L. (2002). *The landscape of history: How historians map the past.* Oxford: Oxford University Press.

Goffman, W. (1964). On relevance as a measure. *Information Storage and Retrieval, 2*(3), 201–203.

Goldman, A. I. (1993). *Philosophical applications of cognitive science.* Boulder, CO: Westview Press.

Grice, P. (1989). *Studies in the way of words.* Cambridge, MA: Harvard University Press.

Habermas, J. (1998). *On the pragmatics of communication* (M. Cooke, Ed.). Cambridge, MA: MIT Press.

Harman, G. (1999). *Reasoning, meaning, and mind.* Oxford: Oxford University Press.

Harter, S. P. (1992). Psychological relevance and information science. *Journal of the American Society for Information Science, 43*(9), 602–615.

Hjørland, B. (1997). *Information seeking and subject representation: An activity-theoretical approach to information science.* Westport, CT: Greenwood Press.

Hjørland, B. (2002). Principia informatica: Foundational theory of information and principles of information services. In H. Bruce, R. Fidel, P. Ingwersen, & P. Vakkari (Eds.),

Emerging frameworks and methods: Proceedings of the Fourth International Conference on Concepts of Library and Information Science (COLIS4) (pp. 109–121). Greenwood Village, CO: Greenwood Press.

Levinas, E. (1969). *Totality and infinity; an essay on exteriority* (A. Lingis, Trans.). Pittsburgh: Duquesne University Press.

Mizzaro, S. (1997). Relevance: The whole history. *Journal of the American Society for Information Science, 48*(9), 810–832.

Pao, M. L. (1989). *Concepts of information retrieval.* Englewood, CO: Libraries Unlimited.

Schamber, L., Eisenberg, M. B., & Nilan, M. S. (1990). A re-examination of relevance: Toward a dynamic, situational definition. *Information Processing & Management, 26*(6), 755–776.

Searle, J. R. (1983). *Intentionality: An essay in the philosophy of mind.* Cambridge: Cambridge University Press.

Sperber, D., & Wilson, D. (1986). *Relevance: Communication and cognition.* Cambridge, MA: Harvard University Press.

Wilson, P. (1973). Situational relevance. *Information Storage & Retrieval, 9*(8), 457–471.

Wittgenstein, L. (1958). *Philosophical investigations* (3rd ed.). (G. E. M. Anscombe, Trans.). New York: Prentice Hall.

http://jodi.ecs.soton.ac.uk/Articles/v04/i04/Soergel/

Journal of Digital Information, Volume 4 Issue 4
Article No. 257, 2004-03-17
To cite this paper please include the details above in the full reference
Themes: Digital libraries, Information discovery

Peer reviewed paper

Reengineering Thesauri for New Applications: the AGROVOC Example

Dagobert Soergel*, Boris Lauser, Anita Liang, Frehiwot Fisseha, Johannes Keizer and Stephen Katz
*College of Library and Information Services, University of Maryland, College Park
Email: dsoergel@umd.edu
Food and Agriculture Organization (FAO) of the United Nations,
Library & Documentation Systems Division, 00100 Rome, Italy
Email: {frehiwot.fisseha; boris.lauser; anita.liang; johannes.keizer; stephen.katz}@fao.org
Key features: References; Table 1, 2, 3, 4, 5, 6; Figure 1

Abstract

Existing classification schemes and thesauri are lacking in well-defined semantics and structural consistency. Empowering end users in searching collections of ever increasing magnitudes with performance far exceeding plain free-text searching (as used in many Web search engines), and developing systems that not only find but also process information for action, requires far more powerful and complex knowledge organization systems (KOSs). The paper presents a conceptual structure and transition procedure to support the shift from a traditional KOS towards a full-fledged and semantically rich KOS. The proposed structure also complies with other interoperability approaches like RDFS and XML in the Web environment. AGROVOC, a traditional thesaurus developed and maintained by the Food and Agriculture Organization (FAO) of the United Nations, serves as a case study for exploring the reengineering of a traditional thesaurus into a fully-fledged ontology. We start the process of developing an inventory of specific relationship types with well-defined semantics for the agricultural domain and explore the rules-as-you-go approach to streamlining the reengineering process.

Contents

- 1 From thesauri to rich ontologies
  - The problem
  - The relationship of traditional KOS to ontologies
  - Potential benefits of future generation KOSs
  - The process of reengineering: The rules-as-you-go approach
- 2 AGROVOC: a multilingual agricultural thesaurus
  - Background
  - Applications and related terminologies
  - Conceptual structure of AGROVOC
  - Semantic problems of AGROVOC
  - The need for reengineering AGROVOC into

1 From thesauri to rich ontologies

1.1 The problem

Empowering end users in searching collections of ever increasing magnitudes with performance far exceeding plain free-text searching (as used in many Web search engines), and developing systems that not only find but also process information for action, require considerably more powerful - and complex - knowledge organization systems (KOS) than the classification schemes and thesauri that currently exist. Such systems must serve the following functions, among others:

- Improved user interaction with the KOS on both the conceptual

and the term level for improved query formulation and subject browsing, and for more user learning about the domain.
- Intelligent behind-the-scenes support for query expansion, both concept expansion and synonym expansion, within one language and across languages.
- Intelligent support for human indexers and automated indexing/categorization systems.
- Support for artificial intelligence and semantic Web applications.

All of these functions require semantic relations that are more expressive and nuanced than the few rudimentary categories and relationships found in traditional thesauri and classifications.

A typical scenario in information retrieval illustrates some of the shortcomings of current free-text search engines such as Google. A farmer is interested in finding out about *rice* and starts a search by entering the string '*rice*'. The results returned in response to the query immediately indicate several problems. First, because the system performs the search based on the actual text string entered rather than on an interpretation of the meaning of the string, many irrelevant results are retrieved. This occurs because the query term itself is ambiguous (i.e. *rice* can refer to the grain, to the university in Houston, or to the name of an author, among others). Further, there are millions of results with no apparently meaningful arrangement. To find something of possible relevance, the user may need to click and scan page after page of the retrieved results. Finally, the user is stuck with the results that have been retrieved; to find other related resources, such as *rice cultivation*, the user must start from the beginning again and formulate a different query, despite the fact that the new query corresponds to concepts related to the original query. The problem becomes evident: The biggest challenge in information retrieval is **concept identification** in a specific **domain of interest**!

an ontology
- 3 Conceptual model: combining thesauri and ontologies
  - The basic model
  - Model extensions
  - Limitations
  - Implementation
  - Related approaches
- 4 The AGROVOC case: exploring conceptual relationships for the agricultural domain
  - The logical generic relationship
  - The part-whole family of relationships
  - Further relationship examples
- 5 Exploring the rules-as-you-go approach for the case of AGROVOC
- 6 Implications and further work
- Acknowledgements
- References
- Appendix: Towards an XML/RDF specification for KOS

In contrast, in a semantics-driven information retrieval system, the system would recognize, i.e. "understand", that the string '*rice*' was ambiguous; it would then request clarification from the user as to which of the possible meanings was intended. Only then, after the user disambiguated the term, would the system execute the search. The system would then retrieve only those resources that had been semantically marked up (through manual or automatic indexing) with the concept of *rice*, no matter what words or even languages are used in the resources to refer to *rice*. Moreover, because the system is semantically rich, it not only presents results that are based on understanding the user's request, it also offers related concepts the user might not have thought of initially. Based on a *<hasPest>* relation, the system could display such concepts as *rice weevil* and *rice moth*. Searching on these latter concepts could in turn lead to concepts on pesticides used on rice, and so on. The system could retrieve not only information directly pertinent to the user's query but also help the user explore and clarify the information need and find useful related information. In this scenario, a KOS has two functions: assisting the user with exploring the topic of the query, and supporting intelligent automatic indexing (metadata assignment) through statistical and syntactic-semantic analysis and "understanding" of text; both functions require a KOS with a rich and precisely defined semantic structure.

To accomplish these and other more sophisticated tasks, the new KOS must marry the conceptual structure of full-fledged ontologies - well-structured hierarchies of concepts connected through a rich network of detailed relations that support concept retrieval and reasoning - with the terminological

richness of good thesauri. While existing KOSs do not provide the full set of precise concept relations needed for reasoning, existing KOSs, both large and small, represent much intellectual capital. This paper explores the question of how this intellectual capital can be put to use in constructing full-fledged KOSs.

1.2 The relationship of traditional KOS to ontologies

Reengineering thesauri, classification schemes, etc., into ontologies means building on the information contained in them and refining that information as needed. Consider the relationships given in the ERIC Thesaurus (ERIC = Educational Resources Information Center) with those given in a hypothetical ontology as shown in Table 1.

Table 1: Statements and rules of a hypothetical ontology versus the information given in the ERIC thesaurus (broader term (BT), related term (RT))

Eric Thesaurus	Hypothetical ontology
reading instruction BT instruction RT reading RT learning standards **reading ability** BT ability RT reading RT perception	**Statements:** **reading instruction** <*isa*> instruction <*hasDomain*> reading < *governedBy*> learning standards **reading ability** <*isa*> ability <*hasDomain*> reading <*supportedBy*> perception
	Rule 1 **Instruction in a domain should consider ability in that domain:** **X *shouldConsider* Y** **IF** X <*isa (type of)*> instruction **AND** X <*hasDomain*> W **AND** Y <*isa*> ability **AND** Y <*hasDomain*> W yields: : The designer of *reading instruction* should also consider *reading ability.* **Rule 2** **X *shouldConsider* Z** **IF** X <*shouldConsider*> Y **AND** Y <*supportedBy*> Z yields: The designer of *reading instruction* should also consider *perception.*

The inferences given rely on the detailed semantic relationships given in the ontology. But the ERIC thesaurus gives only some poorly defined broader term (BT) and related term (RT) relationships. These relationships are not differentiated enough to support inference.

For another example, consider the hypothetical ontological relationships and rules we could

formulate with these relationships in an example taken from the AGROVOC thesaurus (described in detail in section 2) in Table 2.

Table 2: AGROVOC relationships compared with more differentiated relationships of a hypothetical ontology (narrower term (NT), broader term (BT))

AGROVOC	Hypothetical Ontology
Undifferentiated hierarchical relationships in AGROVOC **milk** NT cow milk NT milk fat **cow** NT cow milk **Cheddar cheese** BT cow milk	Differentiated relationships in an ontology **milk** *<includesSpecific>* cow milk *<containsSubstance>* milk fat **cow** *<hasComponent>* cow milk* **Cheddar cheese** *<<madeFrom>* cow milk
	Rule 1 **Part X *<mayContainSubstance>* Substance Y** **IF** Animal W *<hasComponent>* Part X **AND** Animal W *<ingests>* Substance Y **Rule 2** **Food Z *<containsSubstance>* Substance Y** **IF** Food Z *<madeFrom>* Part X **AND** Part X *<containsSubstance>* Substance Y

In the context of food and nutrition it makes eminent sense to consider milk and egg as parts of an animal since their nutritional value and safety depend on the nature of the animal and the feed it ingests just as do skeletal meat and organ meat. This is an example of careful definition of relationships.

From the statements and rules given in the ontology, a system could infer that *Cheddar cheese <containsSubstance>milk fat* and, if cows on a given farm are fed mercury-contaminated feed, that *Cheddar cheese* made from milk from these cows *<mayContainSubstance>mercury*. But the present AGROVOC Thesaurus (described in detail in section 2) gives only narrower term/broadr term (NT/BT) relationships without differentiation.

The **limitations of existing KOS** can be summarized as follows:

- **Lack of conceptual abstraction:** thesauri and other traditional KOSs are collections of terms (generic or domain-specific), ordered in a polyhierarchic lattice structure or a monohierarchic tree structure and interlinked with some very broad and basic relationships. The distinction between a concept (meaning) and its lexicalizations (words) is not made consistently, if at all, in such a system, and as such it does not reflect the ways humans understand the world in terms of meaning and language.
- **Limited semantic coverage**: most thesauri do not differentiate concepts into types (such as

living organism, *substance*, or *process*) and have a very limited set of relationships between concepts, distinguishing only between hierarchical relationships, i.e. NT/BT, and associative relationships, i.e. RT. These very rudimentary relationships are not powerful enough to guide a user in meaningful information discovery on the Web or to support inference. They do not reflect the conceptual relationships that people know and that can be used by a system to suggest concepts for expanding the query or making it more specific. Examples:
    - The relation between *cow* and its part *cow milk* is expressed as NT rather than the more semantically expressive relation *<hasComponent>*, so a user who wants to expand the query hierarchically (search for all concepts narrower than *cow* as well) could not distinguish between searching for all cow parts or searching for all varieties of cow;
    - the relation between *mad cow disease* and the animal it afflicts, *cow*, is expressed using RT instead of the more semantically precise relation *<afflicts>*, so the user could not easily assemble a list of all cow diseases and search for recent occurrences;
    - *mad cow disease* and one of its symptoms *anorexia* would also be related using RT rather than the more semantically expressive relation *<hasSymptom>*.

  The concept relations provided by most thesauri force all relations into the two broad categories, hierarchical and associative. Too often the semantic relationships captured in this way are ambiguous and poorly defined. The generalization/specialization relations defined in most thesauri are not adequately developed to be of use for semantic description and discovery of Web resources. Thus there is a need for a richer and more powerful set of relationships.
- **Lack of consistency**: since the relationships in thesauri lack precise semantics, they are applied inconsistently, both creating ambiguity in the interpretation of the relationships and resulting in an overall internal semantic structure that is irregular and unpredictable. Many of the NT/BT hierarchical relationships could, for example, be resolved to the non-hierarchical RT relationship, and vice versa.
- **Limited automated processing**: traditionally thesauri were designed for indexing and query formulation by people and not for automated processing. The ambiguous semantics that characterizes many thesauri makes them unsuitable for automated processing.

To overcome these limitations and enable more powerful searching and intelligent information processing, especially as such capabilities can be made more widely available through the Web, traditional KOSs must be reengineered into KOSs that contain domain concepts linked through a rich network of well-defined relationships and a rich set of terms identifying these concepts. A concept can be represented by many different terms (words or phrases) in multiple languages. This paper refers to terms as **lexicalizations** of a concept. One term can identify several concepts (homonymy) and one concept can have multiple synonymous terms. A concept is conveyed by all its lexicalizations, the domain it occurs in, and by its **relationships** to other concepts. In addition, valid rules and constraints need to be specified to provide additional generalizations over sets of related concepts and to support inference. These systems must also be converted to machine-processable formats based on Web technologies like XML which tag the vocabularies in a standardized way.

In contrast to traditional KOSs, ontologies provide conceptual abstraction and differentiated relationships. Ontologies specifically separate concepts from lexicalizations and thereby better reflect the structure of human understanding of a domain. In ontologies, the semantics are developed through ensuring that each concept within the domain is uniquely and precisely defined and by specifying elaborated relationships among the concepts. The relationships in an ontology are explicitly named and developed with specification of rules and constraints so that they reflect the context of the domain for which the knowledge is modeled.

Given their more precise and unambiguous semantics, ontologies allow further knowledge to be inferred from the knowledge explicitly represented in the ontology. The new (implicit) knowledge could be derived by applying generalization or transitivity rules, the level of applicability of which is limited in a poorly defined KOS like a traditional thesaurus. This added knowledge in the ontology makes it powerful when employed for intelligent information processing. Although there is a huge

cost involved in moving from thesauri to ontologies, there is an expectation that the added power of consistency, precision, and completeness will be worth the investment even though reliable numbers on the return on investment (ROI) of ontology development are hard to come by.

1.3 Potential benefits of future generation KOSs

For emerging KOSs to satisfy user needs, they must improve both information organization and retrieval in a way that was not possible with traditional KOSs. The following potential benefits are expected from such systems:

- **Unique identifiers and formal semantics:** the explicit definition of concepts and relations in an ontology allows a unique identifier to be assigned to each concept. As each concept and relation is explicitly defined as a unique entity, the ontology lends itself to semantic formalization.
- **Internal consistency:** another benefit of explicit semantics is the achievement of internal structural consistency in the expression of knowledge due to the possibility of applying integrity constraints.
- **Interoperability:** clear semantics enables interoperability among different KOSs since corresponding concepts within different KOSs would have the same unique identifier, irrespective of the actual lexicalizations used to express those concepts. Semantic interoperability promotes sharing and reuse of knowledge.
- **Greater information integration:** interoperability among different KOSs makes it possible for machines to recognize and analyze intended meaning of terms from disparate vocabularies. This is possible by using structured meta-information and formal knowledge description such as agreed-upon metadata schemas, controlled domain vocabularies, and taxonomies. The ability to integrate terminologies from different sources maximizes the value of investment made in the ontology.
- **Inferencing capability:** new KOSs have the potential for expressing knowledge beyond what is present in the structure of the system. Unlike traditional KOSs where both concepts and relations are underspecified and very few, if any, axiomatic rules exist, the facts (concepts and relations) and rules that can be derived from an ontology have the expressive capabilities that allow for reasoning.
- **Automated information processing:** new KOSs create improved potential to discover relevant information from different sources by exploring patterns and filtering information using conceptual connections represented in the ontology. This enables question-answering from one or more databases or, using natural language processing (see next bullet), from text.
- **Natural language processing (NLP) support:** offers the possibility of providing a direct reply to a search question that is expressed in natural language, using the enhanced relationships and semantics in an ontology, instead of only returning a list of relevant documents.
- **Search query understanding:** using NLP and semantic processing, a system can understand a query posed in natural language, determine the concepts involved and, where useful, create a Boolean query.
- **Concept-based search:** an ontology can provide context-aware search capabilities specific to the area of interest.
- **Integrated information search/browse support:** text mining on the Web (Web mining) through meaning-oriented access, dynamic organization of information with the possibility for cross-domain links are feasible with emerging KOSs.
- **Search query expansion:** the enhancement, extension, and disambiguation of user query terms become possible with the addition of enriched domain- and context-specific information.

To be an effective tool to facilitate information categorization, integration and retrieval, ontologies should be multilingual, domain-specific, and cross disciplinary at the same time. For maximum application potential they should be developed in a non-proprietary, application-independent, and

machine-processable format to ensure interoperability among different systems.

1.4 The process of reengineering: the rules-as-you-go approach

Reengineering a thesaurus into an ontology entails refining thesaurus relationships, a laborious process. The steps in the process are:

1. Define the ontology structure
2. Fill in values from one or more legacy KOS to the extent possible
3. Edit manually using an ontology editor:
    1. make existing information more precise
    2. add new information

Step 1 is addressed in section 3, which gives an overall conceptual model at a high level of abstraction, and section 4, which begins the process of defining a set of relationship types for the food and agriculture domain by examining relationships in AGROVOC as to their relationship types.

Step 3 is the most laborious. We have plans to streamline this process by implementing intelligent conversion using a "**rules as you go**" approach. The idea is as follows: The KOS editor watches out for patterns; based on these patterns the editor formulates rules that can be applied immediately to all subsequent similar cases as illustrated in the following:

1. An editor has determined that
   cow NT *cow milk* should become *cow <hasComponent> cow milk*
2. She recognizes that this is an example of the general pattern
   animal <hasComponent> *milk* (or, even more general *animal* <hasComponent> *body part*)
3. Given this pattern, the system can derive automatically
   goat NT *goat milk* should become *goat* <hasComponent> *goat milk*
   since goat is an animal and goat's milk ends with the word milk and thus can be seen to be a type of milk.

To automate this approach even more, we plan to build an inventory of patterns such as *animal <hasComponent> body part*, augmented by an ontology that specifies the concepts of type *animal* (*cow, goat, sheep, horse, chicken*, etc.) and the concepts of type *body part* (*skeletal meat part, liver, bone, milk, egg*, etc.). This information would be drawn from AGROVOC itself and other sources, such as Langual, UMLS, and even WordNet. The system can then detect the applicability of these patterns, at least once it saw one example transformed by an ontology editor. The ontology editors will add to the pattern inventory incrementally.

These patterns are a special type of constraint. Other constraints can be formulated and used to limit the options presented to the human editor as thesaurus relationships are refined. The bases for such constraints are the thesaurus relationships, on the one hand, and the entity types of the concepts involved, on the other. Table 3 shows some examples of constraints based on thesaurus relationships.

Table 3: Some relationship constraints

Thesaurus Relationships	Possible ontology relationships
NT/BT	*<hasMember> / <memberOf>* *<includesSpecific> / <isa>* *<hasComponent> / <componentOf>* *<spatiallyIncludes> / <spatiallyIncludedIn>*

	etc.
RT	*<similarTo>* *<growsIn> / <EnvironmentForGrowing>* *<treatmentFor> / <treatedWith>?* *<hasMember> / <memberOf>* etc. Note that the RT relationship often transforms into relationships that are not symmetric. Note further that in a well constructed thesaurus, an RT should not resolve into an *<isa>* relationship. However, reality shows that the RT relationship has been applied to express this relationship. This can be taken as another proof for the weak definition of relationships in many thesauri.

This inventory will constrain the available choices when manually refining a thesaurus relationship to a more specific ontology relationship. Of course, an authorized ontology editor can override such constraints and thereby update the relationship table. As a relationship has been added or refined the inverse relationship is automatically added or refined.

2 AGROVOC: a multilingual agricultural thesaurus

This section describes the AGROVOC Thesaurus and further illustrates the problem of semantic under-specification.

2.1 Background

AGROVOC is a multilingual, structured, controlled vocabulary/thesaurus designed to cover concepts and terminology in agriculture, forestry, fisheries, food and related domains (e.g. environment). It was developed by the Food and Agriculture Organization (FAO) of the United Nations and the Commission of the European Communities in the early 1980s to describe documents and other information resources in a controlled language for indexing and searching. It contains approximately 16,500 descriptors and 10,000 non-descriptors.

AGROVOC is available online and for download in the five FAO official languages (English, French, Spanish, Chinese and Arabic). It is translated into other national languages such as Czech, Danish, German, Italian, Polish, Portuguese, Slovak and Thai.

2.2 Applications and related terminologies

AGROVOC is used for controlled-vocabulary indexing and searching globally and in various systems throughout the FAO. Systems where AGROVOC is used include:

- the AGRIS/CARIS network, an international information system for indexing and retrieval since 1986, coordinated by FAO;
- domain-specific documentation centers around the world;
- indexers, librarians and translators working in the global food and agriculture sector.

Within FAO:

- Electronic Information Management Services (EIMS)
- WAICENT information finder
- The FAO library catalogue online

AGROVOC coexists as one knowledge organization system next to numerous others in the agricultural domain. Among the most important ones at the FAO are:

- FAO Terminology
- ASFA (Aquatic Sciences and Fisheries Abstracts) Thesaurus
- Fisheries Glossary
- One Fish Glossary

There are a number of other thesauri in the food and agricultural sector, developed by other institutions, such as

- US National Agricultural Library (NAL) Thesaurus,
- Commonwealth Agricultural Bureau Incorporate (CABI) thesaurus,
- Langual thesaurus providing an easily accessible hierarchy with 14 facets.

These thesauri basically all follow the same conceptual structure, which will be discussed in the following section. Nevertheless, we will see that, although all these thesauri use basically the same conceptual model, the information contained in them can differ substantially.

2.3 Conceptual structure of AGROVOC

AGROVOC follows a traditional thesaurus approach. It is a collection of terms, definitions, and term relationships. As is the case with most thesauri, a small, standard, non-adaptable set of relationship types is applied to interlink terms.

2.3.1 Equivalence relationships

USE: Since thesauri have been primarily developed for the purpose of indexing and retrieval, this relationship indicates that any term preceding the USE relation should be replaced, for the purposes of indexing documents and formulating queries, by the term following the USE relation. The relationship usually (but not always) expresses synonymy between two terms.

USED FOR (UF): This is the inverse of USE and indicates that term A is USED FOR term B for indexing purposes.

2.3.2 Hierarchical relationships

Narrower Term (NT): if X is a NT of Y, then X is narrower in some sense than Y. For example, *milk* NT *cow milk, grain* NT *rice*.

Broader Term (BT): if Y is a BT of X, then X is broader than Y; for example *cow milk* BT *milk*, *rice* BT *grain*. BT is the inverse of NT.

Given these rather unspecific definitions, BT and NT relationships can be applied to express generic relations, meronymic relations, instantiations, and many others (see section 4).

2.3.3 Associative relationships

Related Term (RT): the thesaurus conceptual model contains the RT relationship to express any

kind of associative relationship between two terms that is not a hierarchical relationship. This relationship is hence very ambiguous in that it is the default for all other relationships.

Hierarchical (NT, BT) and associative (RT) relationships are relationships between concepts. In the thesaurus, these exist only between descriptors. Following a traditional thesaurus approach, AGROVOC distinguishes between **descriptors** and **non-descriptors** (often referred to imprecisely as preferred terms and non-preferred terms). The rationale behind this is that only a descriptor should be used when referring to the concept (for example, for indexing and retrieval); each descriptor uniquely and unambiguously designates a concept. A non-descriptor must not be used for indexing or retrieval; it is linked through a USE cross-reference to the corresponding descriptor that must be used instead. There are no relationships from one non-descriptor to another.

2.3.4 Scope notes

Many descriptors in AGROVOC have a scope note, which can be a definition of a term, a history note, instructions to the indexer or searcher, or simply a comment. The purpose is to provide the user with more detail about the term and its usage.

2.3.5 Top level structure

Currently AGROVOC has more than 1500 top-level terms, i.e. descriptors which do not have a broader term, making it cumbersome to access the thesaurus from a top-level approach and browse through the hierarchy. Superimposed on AGROVOC is the AGRIS categorization scheme; it has more than 100 top-level categories, ordered in a shallow two-level hierarchy. AGROVOC descriptors are mapped to the second level of AGRIS categories. For example, the AGROVOC descriptor *fish farms* is mapped to the AGRIS category *aquaculture production* which is a subcategory of *fishery and aquaculture*. Thus the AGRIS categorization scheme provides high-level organization for information that has been tagged with AGROVOC descriptors.

2.4 Semantic problems of AGROVOC

Given its minimalist conceptual structure, AGROVOC (as other thesauri) has a number of **semantic flaws**. In the following we will use examples to point out the major drawbacks of the current system and develop the rationale for the shift towards a more powerful, expressive, and unambiguous conceptual model.

2.4.1 Ambiguous descriptor to non-descriptor relationship

In AGROVOC, as indicated, USE/UF covers synonyms and formal variants. In addition, the relation also links quasi-synonyms and very specific narrower terms, which the AGROVOC defines as any of the following:

1. "two concepts considered sufficiently alike to be identified by one descriptor"
2. "a concept and its opposite"
3. "more specific concepts encompassed by one descriptor"

Definition 1 deals with semantically very **closely related**, yet separate concepts (so that the terms designating such concepts would not be true synonyms), such as

famine
UF hunger

Definition 2 involves concepts on **opposite** ends of a scale or otherwise in opposition to each other. (With a few exceptions, the terms designating such concepts are antonyms). Example:

hydrophilicity
UF hydrophobicity

Definition 3 indicates that USE/UF can also express a **hierarchical relationship**, for example:

biological competition
UF interspecific competition
UF intraspecific competition

where the fine distinction between *interspecific competition* and *intraspecific competition* is deemed unnecessary for retrieval and therefore abandoned in favor of the more general category.

2.4.2 Ambiguous hierarchical definitions

The BT/NT relationship used to build up the hierarchy is very ambiguous; it lumps together several different types of relationships as the following examples show:

2.4.2.1 <includesSpecific> relationship (*erythrocytes* are a specific kind of *blood cell*):

blood cells
NT erythrocytes
NT leukocytes

2.4.2.2 <hasComponent> relationship (*blood* contains as a component *blood cells*):

blood
NT blood cells

2.4.2.3 The following example shows clearly the discrepancies between different thesauri that apply the ambiguously defined modeling principles:

AGROVOC and **CABI**:

water
NT ice
NT water vapor
...
NT fresh water
NT drinking water

But **ASFA:**

water
RT ice
RT water vapor
...
NT fresh water

Water vapor and *ice* are *phases of water* while *fresh water* and *drinking water* are *kinds of water*, so in AGROVOC and CABI hierarchical relationships lump together several different semantic relationships. For retrieval this is generally useful (a search for *water* should generally find documents on *ice* as well), but for more differentiated retrieval a user may want to ask for *water in all phases* or for *all kinds of water*. There are many other purposes of semantic processing that need more differentiated relationships. In ASFA the *phase* relationship is treated as a RT, an example of

how grouping relationships may lead to inconsistency. Note, by the way, that neither thesaurus includes the concept *liquid water*, which is logically necessary if *water* means *water in any phase*.

There are many more examples in AGROVOC where the currently used BT/NT relationship is used to describe different relationships. The most obvious ones have been identified and are used in our proposal below in section 4.

2.4.3 Ambiguous associative relationships

Like the BT/NT relationships, the associative RT relationships can be refined into more specific relationships. Some examples are given below.

2.4.3.1 ***<hasMember>*** **relationship** (*Anglophone Africa* <hasMember> *Botswana*)

Anglophone Africa
RT Botswana
RT Gambia
RT Ghana
RT Kenya
RT Lesotho
...

2.4.3.2 ***<causes>*** **relationship** (*bleaching <causes> discoloration*):

bleaching
RT discoloration

2.5 The need for reengineering AGROVOC into an ontology

The examples above indicate clearly the ambiguous nature of the relationships in AGROVOC. With respect to future information retrieval and intelligent processing needs, where it will be necessary to combine different KOS in order to give access to different information systems, it becomes evident that a more rigid structure is required. A reassessment of AGROVOC (as well as other thesauri) to transform its UF, NT, BT, and RT relationships into unambiguously defined relationships and hierarchical order will provide the first step towards solving the problem of ambiguity and inconsistency in information description and retrieval.

3 Conceptual model: combining thesauri and ontologies

This section introduces a conceptual model that provides the necessary structure to create precise semantics to facilitate the transition from traditional thesauri to ontologies. Figure 1 shows the high level conceptual model we propose. Its chief characteristic is a clear separation of the concept level, the term or lexicalization level, and the string level. Present thesauri give a more or less muddled representation of information about concepts and information about terms. The proposed structure allows for a clear separation of concept information and term information. This model owes much to the structure of the UMLS.

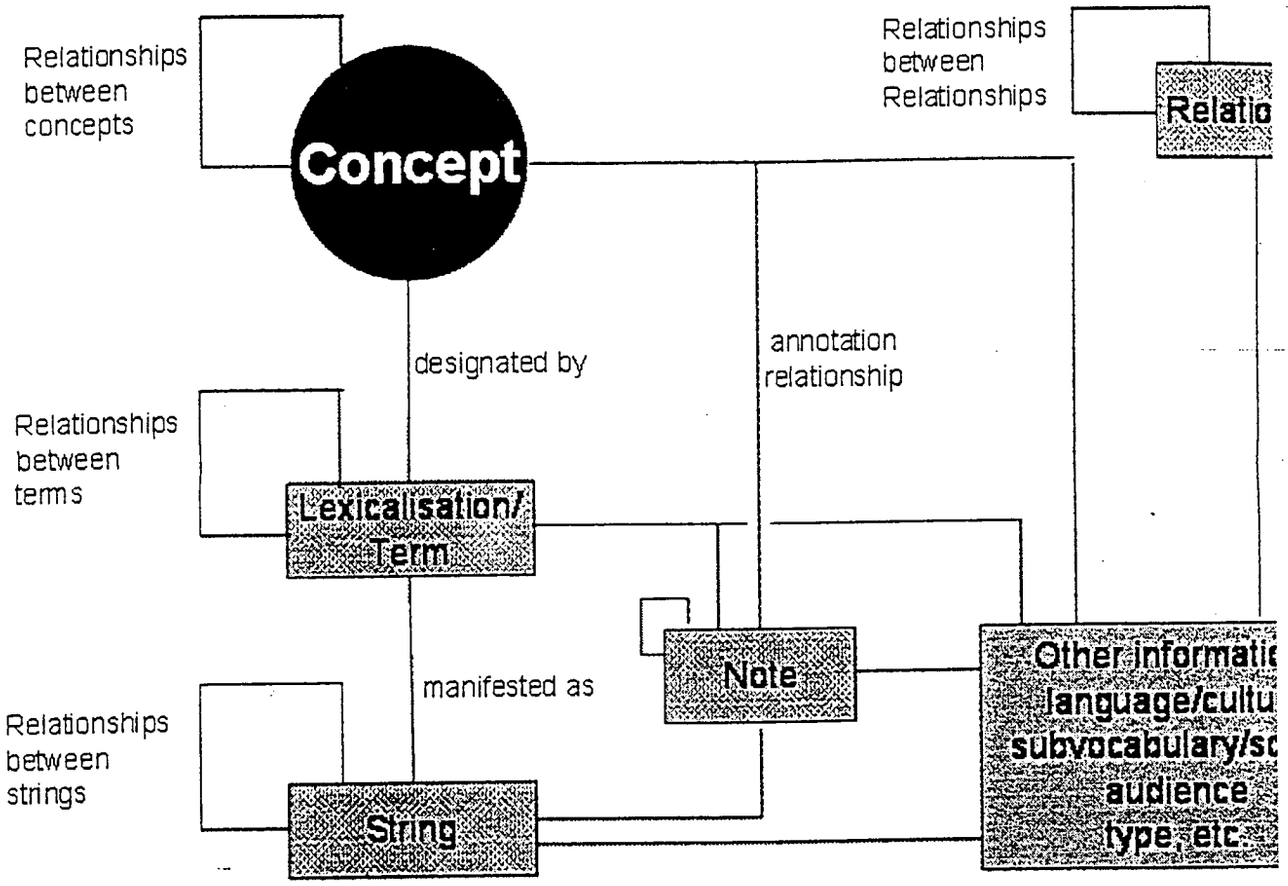


Figure 1. Conceptual model for combining thesauri and ontologies

3.1 The basic model

The following is just the broad outline of the model. Many more types of information could be added. In any event, we consider the model extensible. On the other hand, not all applications will use all features of the model. For example, our model provides for relationships between notes (for example, as hypertext links). This is not possible in all environments but very useful in some. Our intent is to present a framework that can be used for the simplest thesaurus or the most complex and rich ontology in a format that communicates equally to thesaurus and ontology editors with a background in information science, artificial intelligence, or linguistics.

- A **concept** encapsulates meaning.
- A concept can be represented or designated by one or more linguistic expressions, namely **terms** or lexicalizations which can be single words or multi-word phrases (or composite words in agglutinative languages).
- A term, in turn, can take variant forms (singular/plural, variations in case, spelling variants, abbreviations, acronyms); so just as a concept can have many lexical representations, a term can have many **string** manifestations.

Each concept, term, and string can be assigned an identifier, preferably a Unique Resource Identifier (URI); for concepts, UMLS uses Concept Unique Identifiers (CUI), while the Topic Map Standard uses unique subject identifiers. Using unique concept identifiers allows for unambiguous reference to

concepts, as opposed to often ambiguous terms. Concepts can furthermore be assigned notations (such as class numbers in the Dewey Decimal Classification; notations are also called term numbers); notations can be used to maintain a logical, meaningful sequence in hierarchical displays.

Concepts take center stage in our proposed thesaurus/ontology information model; accordingly, relationships between concepts are central. Concepts are arranged in **hierarchies** and have **additional relationships** to other concepts in the network; a hierarchy can be defined on any weak ordering relationship including *isa*, *part-whole*, *spatial containment*, etc. (the relationship must be transitive and not symmetric, but must have an existing inverse relationship, for example *<componenttOf>* is the inverse relationship of *<hasComponent>*). There are many other relationship types, such as *<causes>* ; a scheme of relationship types needs to be defined for the domain of the respective thesaurus. One source for finding relationship types is the detailed analysis of concept relationships present in the thesaurus that is to be reengineered into a richer ontology (see section 4). Each concept should be assigned to an entity type or facet, such as *process*, *function*, *substance*, *living organism* (see, for example, the semantic types in the UMLS Semantic Network); the type of a concept constrains its participation in relationships.

A concept is designated or represented by one or more **lexicalizations** or **terms** in one or more languages; this is the linkage between the concept level and the term level. For examples see Table 4.

Table 4. Concepts, terms, strings (concept and term numbers are fictitious and used only for illustration)

Concept ID	Term ID	Strings manifesting the term
AGROVOC:C316301	AGROVOC: T657210	bovine spongiform encephalopathy, BSE
"	AGROVOC: T657211	mad cow disease, Mad Cow Disease, MCD
"	AGROVOC: T734567	encephalopathy spongiforme bovine, ESB
"	AGROVOC: T734566	maladie de la vache folle, MVF
"	AGROVOC: T700345	encefalopatia espongiforma bovina, EEB
"	AGROVOC: T700346	enfermedad de la vaca loca, EVL
AGROVOC:C014593	AGROVOC: T187953	plow, plows, plough, ploughs, plogh [a frequent misspelling]
"	AGROVOC: T498001	charrue
"	AGROVOC: T498002	materiel de labour

If a term is a homonym (designates more than one concept), several disambiguated terms are introduced. The homonym is linked to each of the disambiguated terms, and each disambiguated term is linked to the corresponding concept. Two terms designating the same concept are called synonyms. Conversely, if one does not agree that concepts per se exist, one can simply view "concept" as a convenient shorthand for an equivalence class of terms that are linked by the *<hasSynonym>* relationship, such as the synsets in Wordnet. A KOS may select a preferred term as the term used to represent the concept or it may make that choice dependent on the audience (for example, veterinarians versus farmers).

Terms can be connected through many **relationships** such as *<hasSynonym>* (with *<hasScientificName>* as a special case), *<hasAntonym>*, *<hasCognate>* (term in a different language from the same root), and *<hasTranslation>*. One might think that the *synonym*

and *translation* relationships are not needed since all terms linked to the same concept would be synonyms or translations. However, two terms may be linked to the same concept yet be used in different contexts, i.e. they are not strict synonyms. If a concept has linked to it several English terms and several French terms, it is not true that just any of the French terms is a good translation for a given English term (see the examples in Table 4). Another example of term-specific relationships is *<hasAntonym>*. For example, *big* and *small* designate opposite concepts but are not antonyms. (The antonym pairs are *big* versus *little* and *large* versus *small*; see Wordnet.)

Finally, a term is manifested in one or more **strings**,as shown in Table 4. Strings can be connected through relationships such as *<hasCaseVariant>*, *<hasSpellingVariant>*, *<hasAbbreviationOrAcronym>*, *<pluralOf>* / *<singularOf>*, which are all subordinate of a broader relationship *<hasStringVariant>*. A term can be seen as a convenient shorthand for an equivalence class of strings that are linked by the *<hasStringVariant>* relationship. A KOS may select a preferred variant as the string used to represent the term or it may make that choice dependent on the audience (as in British versus US spellings). A string, especially an acronym, may belong to several terms, in which case it needs to be disambiguated.

In addition, a concept, a lexicalization/term, a string, or a relationship type can have several types of **notes** (definitions, usage notes, comments, image, etc.) in different languages (in the case of multilingual thesauri). Just like concepts and terms, notes can be related to each other through relationships such as *<hasTranslation>*, *<hasSimplifiedVersion>*, *<hasOtherDefinition>*, or any other type of hyperlink. Many other pieces of information about terms can be added, for example, case frames for verbs (in case the verb has a case frame different from the case frame for the corresponding action concept) or register (see below) or whether the term is the preferred term for the concept. Administrative data will be accommodated as well.

Relationship types themselves can form **relationship hierarchy** (i.e. a relationship of relationships), in which more generic relationships are further up in the hierarchy than more specific relationships, for example, *<componentOf>* is a specific kind of *<partOf>* relationship.

Why define concepts, terms, and strings as separate entity types?

First, each of these entity types takes different types of information. Conceptual relationships and other information are associated with concepts. Linguistic information, such as part of speech and how a term combines with other terms into sentences, usage, or information on etymology, are associated with terms. Information such as that a string is an acronym is associated with terms. Usage information may sometimes be associated with strings; for example, lay people may commonly use a slang abbreviation while professionals use the full string. Definitions are primarily associated with concepts but may also be associated with terms.

Second, this distinction avoids confusion. In a standard thesaurus like AGROVOC, for each concept that is to be used in indexing and searching, a preferred term, and for that term a preferred string, is selected; this string is the descriptor. Non-descriptors are linked only to descriptors, not among themselves. As a result, *BSE*, *mad cow disease*, and *MCD* [which we made for illustration] are all linked to *bovine spongiform encephalitis* as synonyms (or, in some thesauri, as synonym and as abbreviations). But the information that *BSE* belongs with *bovine spongiform encephalitis* and *MCD* with *mad cow disease* is lost. Furthermore, if decisions on terms are made (for example, omitting *mad cow disease* as a non-scientific name), these decisions should apply to all term variants, in the example *MCD* as well.

3.2 Model extensions

As was mentioned above, many more types of information could be added to concepts, terms, strings, notes, and relationships. For example, we might specify an audience (general lay public, K-

12 students by grade level, university students, experts), a subject domain, a scope (as in Topic Maps), or a specially selected subset of concepts and terms to be used for a given application, or all concepts and terms taken from a given source.

Scopes could be defined in many ways. For example, one might define a scope as the conceptual system embedded and expressed in a language (whereas the link from terms and notes to language simply refers to the surface form). Consider the conceptual system underlying Walpiri (an Australian indigenous language); one of its noun classifiers includes *women, fire*, and *dangerous things* (Lakoff 1987). A native speaker of English would find this classifier and the corresponding <*isa*> relationships very curious. Thus one would introduce the category and the <*isa*> relationships with a scope of the Walpiri conceptual system. (By the way, the relationship between these relationships makes sense in the context: fire is dangerous; fire is sometimes started by or anyway related to the sun; the gender of the noun for sun is female). Many such problems, if more subtle, occur in thesauri for international use.

A **subvocabulary** can be extracted using any type information about concepts, terms, strings, and relationships that is available in the thesaurus. Thus one could extract as subvocabulary

- a subset that was selected for a given application;
- all strings that are acronyms (for an acronym reference);
- all scientific names;
- all entries for taxonomic entities for the entire range of living things or for a given large taxon such as insects;
- terms suitable for a given audience.

Each of these subvocabularies provides a specific view on the entire KOS for a given purpose. In online implementations such subvocabularies can be created on the fly or defined as views for certain user groups. But subvocabularies can also be printed or exported (for example, a subvocabulary extracted as the personal KOS of a researcher who maintains an information retrieval system on his or her own computer).

3.3 Limitations

The separation into the concept layer and the term layer is appealing for its simplicity and elegance but it is somewhat of an oversimplification. Terms, particularly terms in different languages, rarely mean exactly the same thing. So the question arises as to when to map two terms to the same concept - and possibly explain shades of meaning and associations in the definition of each term that complements the definition of the concept - and when to create two closely related concepts, possibly under one broader concept. Our model permits any type of relationship between terms. Thus it is possible to introduce conceptually motivated relationships between terms that more accurately reflect the reality of language than the mapping of terms to "concepts". These two representations of conceptual information can coexist within the same system.

3.4 Implementation

All relationships from all layers (concept, term, string) can be stored in the same format within a database. The type of each element should be explicitly given to enable integrity constraints (so that the relationship <*hasSpellingVariant*> is not allowed between two concepts, for example). A concept can be identified by a URI or other number (cleanest solution) or by its preferred term in the base language of the thesaurus (the term being typed as preferred). Likewise, a term can be identified by a URI or other number (cleanest solution) or by the preferred string (the term being typed as preferred). The same holds for strings. The main difference with implementations in most existing thesaurus management software is that relationships between non-descriptors are allowed. Thoughts for an XML/RDF schema for KOS data are presented in the Appendix.

3.5 Related approaches

The proposed conceptual model integrates well with standardization approaches regarding Web technologies like RDFS. The proposed structure shows all aspects of a proposed RDFS-compatible Thesaurus Interchange Format by Matthews *et al.* (2002), which will appear as a W3C note. The proposal is being done in the context of the SWAD-Europe project. The Appendix presents another approach for representing ontology and thesaurus data in XML/RDFS.

4 The AGROVOC case: exploring conceptual relationships in the agricultural domain

The model we introduced has no restrictions on potential relationships to be applied. The model is extensible, and any possible specific relationships can be included. We carried out a preliminary linguistic and conceptual analysis of AGROVOC and found a set of relationships; most of them are well known (but it is important to know that they are needed in the food and agriculture domain), some of them add new nuances. Table 5 lists relationship types found in AGROVOC or otherwise proposed here, and subsections 4.1 - 4.3 give an explanation and examples for some of these relationship types; others appear in examples throughout the paper. This section is not in any way intended as a complete list of relationship types; it merely gives examples to illustrate the additional information and clarity of conceptual structure that can be conveyed through more specific relationships. Much more work, including comparison, is needed to converge on a set of relationships to replace the currently used thesaurus relationships BT, NT, RT, USE and UF in a reengineering of AGROVOC.

Table 5: Concept relationships: Examples

X, Y are concepts

Isa

X <includesSpecific> Y / Y <*isa*> X

X <*inheritsTo*> Y / Y <*inheritsFrom*> X

Holonymy / meronymy (the generic whole-part relationship)

X <*containsSubstance*> Y / Y <*substanceContainedIn*> X

X <*hasIngredient*> Y / Y <*ingredientOf*> X

X <*madeFrom*> Y / Y <*usedToMake*> X

X <*yieldsPortion*> Y / Y <*portionOf*> X

X <*spatiallyIncludes*> Y / Y <*spatiallyIncludedIn*> X

X <*hasComponent*> Y / Y <*componentOf*> X

X <*includesSubprocess*> Y / Y <*subprocessOf*> X

X <*hasMember*> Y / Y <*memberOf*> X

Further relationship examples (some from (Schmitz-Esser 1999)

X <*causes*> Y / Y <*causedBy*> X

X <*instrumentFor*> Y / Y <*performedByInstrument*> X

X <*processFor*> Y / Y <*usesProcess*> X

X <*beneficialFor*> Y / Y <*benefitsFrom*> X

X <*treatmentFor*> Y / Y <*treatedWith*> X

X <*harmfulFor*> Y / Y <*harmedBy*> X

X <*hasPest*> Y / Y <*afflicts*> X

X <*growsIn*> Y / Y <*growthEnvironmentFor*> X
X <*hasProperty*> Y / Y <*propertyOf*> X
X <*hasSymptom*> Y / Y <*indicates*> X
X <*similarTo*> Y / Y <*similarTo*> X
X <*oppositeTo*> Y / Y <*oppositeTo*> X
X <*hasPhase*> Y / Y <*phaseOf*> X
X <*growsIn*> Y / Y <*EnvironmentForGrowing*> X
X <*ingests*> Y / Y <*ingestedBy*> Y

4.1 The logical generic relationship

4.1.1 **X <includesSpecific> Y / Y <*isa*> X** (implies X <*inheritsTo*> Y / Y <*inheritsFrom*> X)

This is the standard generic relationship between concepts. It can be used for hierarchical inheritance (but is not the only relationship for this purpose). Hierarchical inheritance is useful within the thesaurus/ontology to streamline the writing and presentation of definitions and scope notes and for inheriting relationships. Examples (all NT in AGROVOC unless stated otherwise):

chemical soil types <includesSpecific> saline soils

bovine spongiform encephalopathy <includesSpecific> spongiform encephalopathy

cells <includesSpecific> blood cells

blood cells <includesSpecific> leukocytes

leukocytes <includesSpecific> lymphocytes

lymphocytes <includesSpecific> T-lymphocytes UF in AGROVOC

4.2 The part-whole family of relationships

There are several relationships that fall under the part-whole umbrella. Some of these can be displayed in a hierarchical format. However, the direction of the hierarchy, and the direction of hierarchical inheritance, is sometimes from part to whole and sometimes from whole to part. All relationships are shown starting from the whole first.

4.2.1 X <*containsSubstance*> Y / Y <*substanceContainedIn*> X and X <*hasIngredient*> Y / Y <*ingredientOf*> X

Y is the material or substance of which X is made by nature (<*containsSubstance*>) or by man (<*hasIngredient*>). Y loses its identity once it is incorporated into X. There is no implication of an "all and only" relation with respect to the composing substance, where Y is the sole substance making up X; but a subsequent refinement could make this distinction. In AGROVOC, this relationship appears as BT/NT or RT; quite often, it does not appear at all. Some examples are given below:

blood <*containsSubstance*> blood gases

blood <*containsSubstance*> blood lipids

blood <*containsSubstance*> blood proteins

blood <*containsSubstance*> blood cells (borderline case, see comment in 4.2.4)

All NT in AGROVOC

cocoa beverages <*hasIngredient*> cocoa powder

RT in AGROVOC

4.2.2 X <*yieldsPortion*> Y / Y <*portionOf*> X

X <*yieldsPortion*> Y describes a relation between a mass X and a piece Y taken from the mass, for example,

roots <*yieldsPortion*> cuttings RT in AGROVOC

(Note: the example assumes *root cuttings,* but AGROVOC refers to any kind of *cutting.*

4.2.3 X <*spatiallyIncludes*> Y / Y <*spatiallyIncludedIn*> X

This relation is used for objects with spatial extent. It holds when X is an inalienable part of Y, identifiable but not inherently separable. These include body parts and geographical locations.

Asia <*spatiallyIncludes*> East Asia NT in AGROVOC

Transitivity is also a feature of spatial relations. If X <*spatiallyIncludes*> Y and Y <*spatiallyIncludes*> Z, then X <*spatiallyIncludes*> Z.

Asia <*spatiallyIncludes*> East Asia
East Asia <*spatiallyIncludes*> China
→ Asia <*spatiallyIncludes*> China

4.2.4 Y <*hasComponent*> Y / Y <*componentOf*> X

This relationship holds when X is a part of Y that retains its identity as an object even when built into the whole. In addition, each X must be enumerable or nameable, i.e. not part of a mass. Examples:

plough <*hasComponent*> ploughshare RT in AGROVOC

woody plant <*hasComponent*> plant anatomy (i.e. parts of plants) RT in AGROVOC

nucleus <*hasComponent*> chromosome NT in AGROVOC

There are cases when it is hard to decide when to use <*containsSubstance*> and when to use <*hasComponent*>; *blood* <?> *blood cell* is such a borderline case. We decided on *blood* <*hasComponent*> *blood cell* because each blood cell is a recognizable and separate entity. But one could equally strongly argue for *blood* <*containsSubstance*> *blood cell* because a blood cell is part of a mass and cannot be distinguished from others within the mass; it has at most a very weak identity as an object.

4.2.5 X <*includesSubprocess*> Y / Y <*subprocessOf*> X

There are many processes in AGROVOC which could be linked using this relation, for example,

ATP cycle <*includesSubprocess*> phosporylation RT in AGROVOC

4.2.6 X <*hasMember*> Y / Y <*memberOf*> X

X <*hasMember*> Y indicates a relation of membership within a collective or group or organization.

Francophone Africa <*hasMember*> Benin RT in AGROVOC

biotope <*hasMember*> plant not in AGROVOC

pesticide crops <*hasMember*> Artemisia absynthium RT in AGROVOC

Note that transitivity is not a characteristic of membership relations.

4.3 Further relationship examples (some from Schmitz-Esser 1999)

4.3.1 X <*causes*> Y / Y <*causedBy*> X

Examples

overgrazing <*causes*> desertification RT in AGROVOC

Serpulina hyodysenteriae <*causes*> swine dysentery

preharvest sprouting <*causes*> crop losses

4.3.2 X <*instrumentFor*> Y / Y <*performedByInstrument*> X

This relation expresses the fact that concept X is instrumental to achieve, as a result, concept Y. Example:

plough <*instrumentFor*> ploughing RT in AGROVOC

The instrument considered may be one applied by a living being (man, animal) or a machine or a system. The sense of the relation points to the result achieved by the use of the instrument.

4.3.3 X <*processFor*> Y / Y <*usesProcess*> X

This is a case where X indicates a process involved in Y. Examples:

soil injection <*processFor*> fertilization RT in AGROVOC

gonadectomy <*processFor*> sterilization [of organisms] BT in AGROVOC

4.3.4 X <*beneficialFor*> Y / Y <*benefitsFrom*> X

fertilization <*beneficialFor*> crop yield Not found in AGROVOC

4.3.5 X <*treatmentFor*> Y / Y <*treatedWith*> X

Pentosan polysulphate <*treatmentFor*> bovine spongiform encephalopathy Not in

AGROVOC

4.3.6 X *<harmfulFor>* Y / Y *<harmedBy>* X

Found in AGROVOC only indirectly. For example, from:

preharvest sprouting *<causes>* crop losses

one can conclude

preharvest sprouting *<harmfulFor>* crop yield

4.3.7 X *<growsIn>* Y / Y *<growthEnvironmentFor>* X

Halophytes *<growsIn>* saline soils RT in AGROVOC

4.3.8 X *<hasProperty>* Y / Y *<propertyOf>* X

fertilization *<hasProperty>* application rate RT in AGROVOC

blood circulation *<hasProperty>* blood pressure NT in AGROVOC

4.3.9 X *<similarTo>* Y / Y *<similarTo>* X

bovine spongiform encephalopathy *<similarTo>* Creutzfeld-Jakob syndrome RT in AGROVOC

4.3.10 X *<oppositeTo>* Y / Y *<oppositeTo>* X

crop losses *<oppositeTo>* crop yield Not in AGROVOC

4.3.11 Concluding comment

This preliminary analysis shows that AGROVOC contains many relationships that can be made more specific in a reengineering project but that some relationship types are missing so relationships need to be added from scratch. Also it is hardly possible to compile a complete inventory of relationship types. Thus, a generic relationship *Related Term* should still be kept to express residual relationships; such relationship instances should be accompanied by a note that specifies the meaning or intention of the relationship. This will facilitate a later deduction of new relationship types from these residual relationships.

5 Exploring the rules-as-you-go approach for the case of AGROVOC

We explored the applicability of the rules-as-you-go approach to transforming AGROVOC into an ontology. We looked for examples of patterns that could be used. The results are shown in Table 6.

Table 6: Examples for the rules-as-you-go approach

Pattern: plant *<growsIn>* soil type
Rice RT moist soil → rice *<growsIn>* moist soil

Pattern: geographical entity <*spatiallyIncludedIn*> geographical entity

Benin BT West Africa → Benin <*spatiallyIncludedIn*> West Africa

Pattern: geographical entity <*isa*> geographical entity

Benin RT Francophone country → Benin <*isa*> Francophone country

Pattern: body part <*containsSubstance*> (substance | small particle)

blood NT blood gas → blood < *containsSubstance* > blood gas

blood NT blood cell → blood < *containsSubstance* > blood cell

From this exploration it appears that the approach is promising. On the other hand, the rules-as-you-go approach is not error-free; results must be checked by an ontology editor. In some cases, it may not be possible to define a rule.

Another difficulty is illustrated by the concept of *Francophone Africa*. Does this term refer to the set of Francophone African countries, in which case it refers to a *type of geographical entity*, or does it refer to the area (not necessarily contiguous) covered by these countries, in which case it refers to a *geographical entity* (just as *West Africa*).

UF (Used For) may refer to a synonymous or quasi-synonymous term or to a narrower concept (generally *includesSpecific*). Since the terms on both sides of the UF most often refer to concepts that have the same entity type, it is difficult to formulate a rule. One might use an external knowledge source, such as a large dictionary, to detect synonymity and treat the other cases as *includesSpecific*, always subject to editor verification.

The rules-as-you-go approach implies that the reengineering effort should start with the top-most concepts so that entity types and patterns can be detected early.

6 Implications and further work

On the basis of the ideas presented in this paper, we plan to undertake the reengineering of AGROVOC into an ontology system that will also serve the functions of a traditional thesaurus. This involves creating a complete inventory of domain-relevant entity types and relationship types, which we will base on further analysis of AGROVOC and related vocabularies and on existing inventories, such as the UMLS Semantic Network. We tend towards using a frame representation of relationships, as presented in Slaughter and Soergel (2003).

Based on the encouraging results of our exploration of the applicability of the rules-as-you-go approach to AGROVOC, we plan to develop a system that streamlines the reengineering process. We expect that this system will drastically reduce the effort required and thus make the reengineering effort feasible.

The resulting knowledge base will enable

- improved user interaction with the vocabulary for improved query formulation and more user learning about the domain;
- intelligent behind-the-scenes support for query expansion;
- intelligent support for human indexers and automated indexing/categorization systems;

- support for artificial intelligence and semantic Web applications for agriculture, food processing, and food safety.

The approach will be incremental, starting with providing a proof of concept through pilot application that demonstrates that the expected benefits will in fact result. This will provide the basis for planning and efficient implementation of the large-scale effort to reengineer AGROVOC into a full-fledged ontology. We hope this effort can be carried out in a cooperative, distributed, and coordinated environment to promote active participation and ultimately use of a widely accepted Knowledge Organization System for the domain of food and agriculture.

Acknowledgements

We particularly thank all the participants of an expert workshop on the AGROVOC thesaurus and especially all the users and developers for their valuable feedback provided over the years. Their feedback contributed in great part to this proposal. We also thank the anonymous reviewers for their helpful comments.

References

Lakoff, George (1987) *Women, fire, and dangerous things* (University of Chicago Press)

Links to RDF and XML thesaurus formats and related topics
http://www.w3c.rl.ac.uk/SWAD/thes_links.htm

Matthews, B.M., Miller, K. and Wilson, M.D. (2002) "A Thesaurus Interchange Format in RDF". Submitted to the *Semantic Web Conference 2002*

Schmitz-Esser, W. (1999) "Thesaurus and Beyond: An Advanced Formula for Linguistic Engineering and Information Retrieval". *Knowledge Organization*, Vol. 26, No. 1, 10-22

Schmitz-Esser, W. (1999) "Gedankenraumreisen - neue Thesaurusstrukturen, multimedial präsentiert, machen Anregung, Spielen, Lernen, Finden möglich für jedermann". In *Proc. DGI-Jahrestagung*, Hamburg (Frankfurt/M.: DGI), pp. 347-353

Slaughter, Laura and Soergel, Dagobert (2003) "How physicians' answers relate to health consumers' questions". *Proc. 66th ASIST Annual Meeting*, Long Beach, CA, Oct. 19-22 (Medford, NJ: Information Today), pp. 28-39

Appendix: Towards an XML/RDF specification for KOS

Proposals and schemes for encoding KOS data range from general schemes, such as RDF (with the OWL extensions) and the Topic Map standard to much simpler and specific schemes for encoding thesaurus data as specified in ISO 2788, which inherit all the limitations of this standard and are of equally limited usefulness for the new tasks ahead. Something in between is needed. After some preliminary remarks on the functions of such a standard, we make a proposal that is parsimonious but allows for encoding very rich data. For this reason, it may seem a bit opaque at first. The first author will be happy to entertain any questions or comments.

Functions to be served by standards for machine-readable KOS

1. Input of KOS data into programs / transfer of thesaurus data from one program into another

1.1 Format for original input files (but XML difficult for that, use a more user-friendly format, such as inputting a hierarchy with levels specified by the number of dots at the beginning of a line)
1.2 Transfer from one KOS development program to another
1.3 Transfer from a KOS development program to an information system that uses a KOS for authority control, query expansion (synonym and /or hierarchic), display/browse/search, or other purposes
1.4 Transfer from a KOS development program to a KOS display / browse / search program

2. Querying KOSs and viewing results (for example, using Z39.50)

2.1 By people
2.2 By systems to use data from external KOSs for query term expansion, etc.

3. Identifying specific terms/concepts in specific KOSs

This requires rules for URIs that uniquely identify specific term/concept records in specific KOSs. Probably requires some sort of name resolution service (such a KOS registry)

3.1 Links from one KOS to another
3.2 Indexing terms/concepts in the metadata for an object, or any other reference to a term/concept in a text/object

Elements of an XML KOS data specification

This schema is parsimonious yet allows the recording of many types of data. It gives enough information to derive a full XML specification.

This specification assumes that data from each source are grouped, so that source attribution is not needed for each element; otherwise the structure would be much more complex. This works for a communications format but not for an internal database format.

The term itself is indicated in a relationship of type TERM. This allows for terms in multiple languages for the same concept and simplifies the schema since elements in term would be the same as in relationship target.

Addition of the scope element was inspired by the Topic Map Standard.

Most schemes advanced for KOS data hard-code the permissible relationship types as tags. This makes it very hard to introduce new relationship types. The scheme proposed here is based on a more elegant principle: it simply provides a generic syntax for recording relationships and makes the relationship type a data element recorded in an appropriate tag. The scheme needs a method (not given here) for indicating a relationship set defined elsewhere and used within the source or for defining a relationship set for the source. A relationship must specify the relationship types and domain and range for each. RDFS could be used for this specification (RDF object classes are entities, RDF properties are relationship types). A system processing data organized in this scheme would process a relationship instance as follows: look up the relationship type in the relationship set and get the proper entity type for domain and range, respectively. Then check the entity type in the source (domain) slot and the target (range) slot to verify agreement with the restrictions. The scheme is neutral as to how concepts, terms, and relationship types are identified. The identifiers could be URIs, system-assigned numbers, character strings representing codes or character strings representing terms.

Default is minOccurs="1" maxOccurs="1"

Source (minOccurs="0" maxOccurs="unbounded")

Pointer to or definition of relationship set used

Unit: Concept or term or group of terms (minOccurs="0" maxOccurs="unbounded")

Unique identifier

Type of concept/term [from list of values, to include facetHead]

Hierarchy position (minOccurs="0" maxOccurs="unbounded")

Hierarchical level

Class number / notation

Scope for which this concept/term holds (minOccurs="0" maxOccurs="unbounded")

Relationship (minOccurs="0" maxOccurs="unbounded")

Relationship type

Relationship target

/* See below for structure. */

Relationship strength (minOccurs="0" maxOccurs="1")

Audience level /* Of this relationship */ (minOccurs="0" maxOccurs="unbounded")

Perspective /* Of this relationship */ (minOccurs="0" maxOccurs="unbounded")

Scope for which this relationship holds (minOccurs="0" maxOccurs="unbounded")

Relationship, added information (minOccurs="0" maxOccurs="unbounded")

/* This could be a scope note explaining the relationship, an image illustrating the relationship, another term, etc. */

Type of added information /* Relationship types might be reused here. */

Relationship target

Audience level /* Of this piece of info. */ (minOccurs="0" maxOccurs="unbounded")

Perspective /* Of this piece of information */ (minOccurs="0" maxOccurs="unbounded")

Where relationship target has this structure (unifying term, text, images, multimedia document)

Relationship target

Type

/* Includes types of terms (descriptor, other preferred term, non-preferred term and types of texts and other documents, may be an elaborate hierarchy. */

Target value (a term or a document)

Term

Term variant (minOccurs="0" maxOccurs="unbounded")

Type of variant

/* Such as Preferred Spelling, other SPelling, ABbreviation, Full Term. */

Term form (complete term or Stem plus suffix)

Complete term

Stem plus suffix

Stem

Suffix

Document

Language (zero to many, exactly one for terms)

Audience level /* Of this relationship target */ (minOccurs="0" maxOccurs="unbounded")

Perspective /* Of this relationship target */ (minOccurs="0" maxOccurs="unbounded")

Scope for which this/term holds (minOccurs="0" maxOccurs="unbounded")

Natural Language Thesaurus: A Survey of Student Research Skills and Research Tool Preferences

VICTORIA REDFERN

ABSTRACT This paper reports the results of a University of Canberra Library survey of student research knowledge, skills, tools and resources. Students are experiencing difficulties interrogating databases, the internet and library catalogues because of the lack of consistency in terminology and various methods of interrogation. This research was an investigation of the need for an online contemporary natural language thesaurus. This report provides evidence that students interviewed in this survey support the idea and it is recommended that further research on a semantic contemporary natural language thesaurus be conducted and focused on higher education in Australia.

Over the past ten years there has been an increase in various research tools such as databases, online journals, and the internet. Students experience difficulties using databases, library catalogues, online journals and the internet because of the lack of consistency in terminology, recognition of appropriate search terms and various methods of interrogation and this creates an inefficient and time wasting experience. In order to find better ways of helping students, the University of Canberra Library conducted a pilot study to investigate whether an online contemporary natural language thesaurus would assist students with their research. 'Natural language' is 'any naturally evolved language as opposed to artificial languages constructed.'[1] Accordingly, in order to evaluate the need for the contemporary natural language thesaurus, it was first necessary to evaluate student knowledge of library and database terminologies and research tool usage and skill.

Student Research Strategies

The keys to successful research are effective research strategies and a willingness to learn to explore all available possibilities in order to obtain reliable and comprehensive results.

Before commencing research it is essential for students to understand the objective, decide what subject and search headings cover the topic and to use a variety of resources effectively.[2] Accordingly, to find material in electronic databases, online journals, library catalogues and the internet, it is imperative to formulate appropriate search terms using key words and phrases. Controlled vocabulary searching and free-text searching are different methods of

Victoria Redfern, Student in Master of Educational Leadership, University of Canberra Library. Email: victoria.redfern@canberra.edu.au

vocabulary searching and free-text searching are different methods of interrogation. A controlled vocabulary search interrogates a database by using words that exist in the built-in taxonomy of that database. A free-text search interrogates a database by using any word to find incidences of that word occurring in the database.

There are recent research technologies being developed such as natural language thesauri, and semantic searching. The natural language thesauri interrogates on terms based on commonly accepted speech vocabularies. Semantic searching interrogates on terms semantically linked to find associated materials on the internet as well as databases and other electronically authored resources.

The Library of Congress, database developers, and web page authors use different terminology. Consequently, students have to become conversant with different interrogation methods such as controlled vocabulary searching or free-text searching and the varying terminologies of research tools and databases.

Defining the research question, establishing search terms, definitions and descriptors and using these terms for interrogating the library catalogue, electronic databases, online journals and the internet is complicated by the differences in the thesauri, truncation, Boolean operators and wild cards as well as subject headings. Therefore, to research one topic researchers have to try different thesaurus terms and keywords according to the design of the database or library catalogue. Consequently, because of the differences between databases and their content, varying outcomes using the same search term will result.

Feitzer[3] says that traditional description and indexing is stretched to the extent that the concept of the 'main entry' has been replaced by a multitude of contributors such as database developers and authors who share responsibility for information and this in turn results in inconsistencies in terminology and the interrogation techniques required. Therefore, for consistency, a semantic webbed thesaurus using natural language is needed to enable interrogation across boundaries previously constructed by cataloguers, database developers and web page authors. This would help researchers by providing a simpler interrogation tool and enhance research methodology. Consequently, this leads to gaining required literature and information quickly. In order to evaluate if further research should be undertaken it was essential to conduct a survey to evaluate current student research skills and tools used.

A similar study to evaluate student research skills and whether the internet was found to be beneficial or frustrating was conducted by Lindsay and McLaren[4] at the University of Strathclyde in England. The survey focused on

35 participants who were studying for their Bachelor of Education in Design and Technology.

Before the University of Canberra Library survey two major problems that students had experienced were identified. These problems were that library and database terminology and differing interrogation techniques are required because of the design of library online catalogues, electronic databases and the internet.

Language, speech, terminology and ontology are constantly evolving. Database thesauri, research descriptors and Library of Congress classifications and subject headings are constantly added to, altered and deleted according to language and vocabulary accepted as natural language and being in common use. Consequently, because of the differences between Library of Congress subject headings, database designs, search terms, keywords and interrogation methodologies, there is a lack of consistency.

An example of inconsistency is shown in research term variations. The example is the contemporary natural language term, 'flight delay'. 'Flight delay' is not a Library of Congress subject heading. However, Academic Search Elite refers the user to 'Avigation Easements' whilst also having a spelling error. Additionally, acronyms such as 'SARS' are accepted as contemporary natural language but alterations and additions to the Library of Congress Subject Headings can take a long time to appear in directories.

Table 1
Research Term Variations

Research Term	Library of Congress Classifications 'Subject Headings'	Academic Search Elite 'Subject Terms'
Flight delay	No	See Avigation (sic) Easements
Day dreaming	No	Use Fantasy
Debt management	No	Yes
Fire back-drafts	No	No
Flexible workplace practices	No	No
SARS	No	Yes
SMS	No	Use Text Messages

There are research terms such as 'fire back-drafts' and 'flexible workplace practices' in existence as contemporary natural language that are not classified in the Library of Congress subject headings or as subject terms in databases such as Academic Search Elite.

Besides typographical errors such as 'Avigation', other examples of differences are: Ingenta links to Merriam Webster Collegiate Dictionary for keywords, Expanded Academic has no subject heading list but matches words in articles (free-text), some databases have Boolean operators and others operate on wild cards.

Natural Language Thesaurus

The Library of Congress designs their data elements in a standardised subject heading format. Database developers such as Expanded Academic design their elements interrogated by free-text querying. Field[5] supported natural language research results because adult educators are paying more attention to those students who are from varying ethnic backgrounds and those for whom English is not their natural tongue. Field sees one way of helping these students through new forms of communication media is through tools such as the Online Thesaurus of Natural Language.

Knapp, Cohen and Juedes[6] conducted research on the use of natural language interrogation to determine the potential value to the humanities and social sciences. This research developed because the humanities cross many subjects and many synonyms may be used to cover one concept. They found that even though databases may be interrogated using controlled vocabularies and free-text, there was significant evidence that free-text produced unsuccessful search results because of 'the inability of the searcher to think of all the terms an author may have used'. Knapp, Cohen and Juedes[7] also found that the use of natural language querying did increase the number of relevant citations and this included databases with controlled vocabularies. Additionally, they add that as more full-text databases have become available and there is increased growth in the internet, natural language searching has become 'increasingly vital to ensure retrieval of relevant information'. This being so, to develop a contemporary natural language thesaurus that covers all academic disciplines would be time and resource expensive. Naturally, it would be difficult to develop the thesaurus to cover all resources. However, with the increasing use of digital tools, by concentrating only on electronic formats and constructing the thesaurus one discipline at a time, such as the humanities and social sciences initiative of Knapp, Cohen and Juedes,[8] it may be possible to develop an online contemporary natural language thesaurus.

Methodology

Academic research not only enhances and promotes the knowledge base of available information, it also provides opportunities to learn and develop research methodologies, skills and excellence. The University of Canberra

Library survey was designed firstly to evaluate student knowledge and skills in using different search terms, keywords, descriptors and subject headings for the library catalogue and databases. Secondly, it was to evaluate student use of research tools and research methodologies. The final, and major question of the survey was to establish that if a contemporary natural language thesaurus was available and linked to library catalogues, databases and web pages, would students think this would assist their research and why.

Before the survey commenced, the following limitations were noted: the survey was not expected to represent all students studying at the university as there are a number who are studying by distance education, flexible delivery and online via WebCT and additionally, some students do not attend the library.

The purpose of the University of Canberra College is to prepare students for university studies. Their students were excluded as there was the assumption that they were not yet studying at the same level as undergraduate, postgraduate or masters level students and therefore may not represent the overall cross section of library users. It was expected that some students would not be interested in the survey. It was hoped that the survey would obtain a sample representative of cultural background, gender and age, and from both undergraduate and postgraduate respondents.

The survey was held in the University of Canberra Library foyer during weeks 7-10 of Semester 1, 2004. The study was both quantitative and qualitative and data were collected through questioning and through numerical sequencing.

The interview used ethnographic semi-structured questions and questions that utilised numeric ranking. This was to better understand student knowledge, to extrapolate further information and to allow participants to add any additional comment or opinion. Participants approached were of any nationality, gender or age. The survey was conducted during weekdays, weekends and evenings in order to gain a representative sample. Participants interviewed were undergraduate, postgraduate or masters level students. Twenty students were interviewed. Initially it was intended that at least 35 participants would be interviewed but because a consistent and definite pattern was found it was decided the number would be reduced to 20. Each interview was planned for 10 minutes but this was flexible and often increased for participants who wanted to provide additional input for the contemporary natural language thesaurus question.

The survey instruments were in four sections: Student Demographics, Research Terminology Skills, Research Tool Preferences and Online Thesaurus of Natural Language Search Terms.

Student Demographics

Student Demographic questions were designed to establish the student level of experience. There were four questions. The first was to confirm that they were enrolled at the University of Canberra, the second was their full time or part time status, and the third and fourth were to establish how long they had studied at UC and to establish past research experience.

Result

Although the participants were randomly selected, fewer undergraduate students were represented than had been expected. Those who were included and in their first semester at the University of Canberra had previous experience at CIT, technical college, and the University of Canberra College or had previously dropped out during their first undergraduate semester at another university.

Research Terminology Skills

Research terminology skill questions were designed to establish student understanding of basic library and database terminology as well as gaining information on their research methodology.

In order to ease the participant into the interview three questions were asked. These questions were: 'How do you go about starting an assignment?' 'Why do you do it this way?' and, 'What have you learned by doing it this way?' These three questions provided the primary answers that they first look for keywords and phrases and then follow on with searching on the internet and conferring with fellow students. Because participants being interviewed were now thinking about search terms, this was used as an introduction to the next set of questions on research terminology skills.

Research terminology skills comprised four questions that required the participant to correctly identify the terms 'search term', 'subject heading', 'descriptor' and 'keyword'.

'Search term' was defined as 'A word or phrase input by the user to find those records on the database that contain that term';[9] 'Subject heading' was defined as 'The word or group of words under which books and other material on a subject are entered in a catalogue in which the entries are arranged in alphabetical order';[10] 'Descriptor' is 'An elementary term used to identify a subject';[11] and 'Keyword'[12] is defined as 'Grammatical element which conveys the significant meaning in a document'. Although these terms were used as a basis to define meaning, the participants were not expected produce these definitions as their answer. The intention of the question was to gain an indication whether they knew where the term is used and how they can relate that term to their research.

The percentage figures when averaged mean that, of the total 20 participants, 52% understood the terms and 48% did not understand the terms.

Table 2
Student Knowledge of Terminology

Term	Did Not Know Term		Unsure		Vaguely Sure		Pretty Sure		Knew Term	
	n	%	n	%	n	%	n	%	n	%
Search Term	8	40	4	20	2	10	2	10	4	20
Subject Heading	4	20	2	10	3	15	5	25	6	30
Descriptor	16	80	1	5	1	5	2	10	0	0
Keyword	0	0	0	0	1	5	2	10	7	85

Note: The percentage in each row is the percentage of all participants who understood the term for that row.

Locating Search Terms

Locating search terms consisted of a list of 12 items from which the participant selected those they use. Participants were asked to nominate the tool or tools that they have used for locating search terms. They were asked the same question for each tool. The question for print format tools was: 'Do you look in a thesaurus on the library shelves or somewhere else to find your search terms?' The question for digital format tools was: 'Do you use the electronic databases on computers such as here in the library, at home or work to find your search terms?' The questions were asked the same way for the print format and digital tools. The answers were noted on a survey sheet by the interviewer.

The three most frequently mentioned tools used by participants for finding search terms are:

Internet (Google or similar web search tool)	70%
Ask other students	60%
Electronic databases	50%
Equal fourth are:	
Dictionary of research terms	45%
Online journals and serials	45%
Browse the library shelves	45%

Additionally, in ascending order, the six least used for finding search terms are:

Newspapers	1 student
Printed journals	1 student
Television documentaries	1 student
Index of text books	3 students
Directory of research terms	3 students
Thesaurus	3 students

Table 3
Locating Search Terms

Tool	Total n=20	
	n	%
Thesaurus	3	15
Directory of research terms	3	15
Dictionary of research terms	9	45
Electronic databases	10	50
Online journals and serials	9	45
Online library catalogue	5	25
Browse the library shelves for ideas	9	45
Internet (Google or similar web search tool)	14	7
Other university web sites	6	30
Ask library staff	8	40
Ask lecturer/tutor	9	45
Ask other students	12	60
Newspapers	1	5
Index of text books	2	10
Printed journals	1	5
Television documentaries	1	5

Note: The above list of tools for locating search terms is in the same order as the participants were interviewed.

Research Tool Preferences

Research tool preferences investigated what research tools students use.

The analysis showed that all of the participants surveyed begin their research in the same way. After looking for keywords or phrases in their research question they then refer to their prescribed texts and other books for related materials on their topic. They then search the library catalogue,

databases and the internet and gather the materials together. This was a consistent theme of all participants. A second recurring theme the participants revealed was that their main research tools are digitally based, such as databases and the internet. Additionally, before undertaking their research they consistently attempted to obtain textbooks in the library, however, because these were often not available participants would then turn to digital resources.

The question was asked: 'What tools do you use most frequently for your research?' Participants were shown the research tool list and asked to place them in order of importance and advised they could add additional tools to the list.

Table 4
Research Tool Preferences

Tool	Importance Order	Total Respondents n=20 %
Internet	1	65
Library books	2	50
Online serials	3	45
Electronic databases	4	35
Others – Web CT	5	10
Printed newspapers and journals	6	5
Other libraries	7	5

The tool most often used is the internet (65%) and 'Other Libraries' came last (5%).

Two participants who were studying law said they only accessed books in the library and the law databases. Two were studying Landscape Architecture and said they did not access databases because the majority of their work was visual. Conversely, the Lindsay and McLaren[13] study revealed that although their participants were studying design they still used the internet for benchmarking, market research and consumer appraisal. Therefore, the discipline being studied may not have a bearing on the tools that participants use because many academic disciplines cross over into other areas.

Some participants said that the major benefit of online tools such as databases was that information was up-to-date. Consequently, because it takes longer for books to be published, there was no guarantee that libraries would add them to their collection. Digitally available research findings were more readily available than library books. Conversely, some participants preferred to

use books because they were more comprehensive than electronic materials and they also perceive that to do research without books was non-academic. Two participants said that because they are tactile people, they learn better from books and find them easier to handle than computer keyboards. Overall, most participants felt that obtaining books is often difficult and when they are available they are often out of date. One participant said that obtaining a pass grade would be difficult without tools such as databases, the library catalogue and the internet because these provide a broader knowledge base. A second participant preferred the internet because of its flexibility and greater amount of information. A third participant preferred the internet and databases for their 'time saving and resource convenience'.

Table 5
Hours per Week spent using Research Tools

Tool	Total Users	Total Hours
In the library using books and journals	18	91 hours
Electronic databases	16	37 hours
Internet	19	200 hours

Two students did not use the library at all, four did not use electronic databases and one did not use the internet.

The Online Thesaurus of Contemporary Natural Language Search Terms

The final question of the survey was to establish if students felt an online thesaurus of contemporary natural language search terms would assist their research. This question was designed to find out whether further research into this was warranted. The question was asked using terms that are not listed in the Library of Congress Subject Headings[14] nor The Contemporary Thesaurus of Search Terms and Synonyms.[15]

The concept of an online contemporary natural language thesaurus linked to research materials on the library web page was explained to each participant.

Participants were asked: 'If search terms were more like the words we use in everyday conversation, such as 'day dreaming', 'debt management', 'fire back-drafts', 'flexible workplace practices' and 'flight delays', and these were available in an online contemporary natural language thesaurus, do you think this would make your research task easier, and why?'

Eighty percent of the participants said 'Yes' and 20% said 'No'.

The two who said 'No' said that if students study in an academic environment they must use academic language as the rigour and discipline of

academic research could otherwise be compromised. Another two students said that the cost of establishing the thesaurus and putting it to use could not be justified. However these same four also said that for students for whom English was their second language, a contemporary natural language thesaurus would be a good idea because it would not only assist them with finding search terms but it would also help them with their English. Of the 80% who said 'Yes', 37% of these offered the opinion that even though they may not find it useful, it could benefit students when their English language skill is not strong. Eighty percent of participants said that contemporary natural language was preferable to library, database and internet 'jargon'.

Discussion

The University of Canberra Library survey has produced some interesting results in all three sections of student demographics, research terminology skills and research tool preferences. The results of the final question about the online thesaurus of contemporary natural language search terms are surprising.

Student demographic results were unexpected in that there was a very strong representation of participants who had studied previously at tertiary level and a number of these were currently studying at postgraduate level. It is therefore of concern that given the sample and the amount of experience in research, a large number of participants observed during the interview appeared uncomfortable and also stated that they were embarrassed that they were unfamiliar with basic terminologies and had trouble establishing search terms. However, even though a number of participants did not know the library and database 'jargon', they still believed they knew how to commence research. Most participants said that frequently their search of databases, the library catalogue and the internet produced many results that were unrelated and useless for their studies.

Research terminology skills analysis has revealed some interesting statistics that display student lack of understanding of library and database search terms and it is surprising that knowledge of terminology was not higher. This suggests that library research skills training should focus more on terminology and teach students how to develop or find search terms before they learn how to interrogate databases, the internet and the library catalogue.

Interestingly, all participants used a variety of tools for locating search terms. It is also noted that one student nominated newspapers and a second student nominated television documentaries for locating search terms. Of further interest is that the usage of the traditional methodology of using Library of Congress subject headings and discipline-based directories to locate search terms was not popular.

Even though the majority of students use digital tools for locating search terms, when they do begin searching for materials their preferences turn firstly to the internet and then library books. Whilst students consider library books are often not available or out of date, they still see them as being more comprehensive than the electronic databases. The most interesting statistic is that 65% of participants prefer to use the internet for their research. Do we really know if the internet provides students with the materials they need, and that they find it accurate, authentic and easy to use?

The amount of time students spend using the library, electronic databases and the internet is not important as long as that time is spent locating quality, accurate and authentic results. However, it is also possible that students are not finding the materials they need. This may be because of their interrogation techniques. Could this be because they are familiar with one search methodology rather than another, and therefore use only one database or search engine and limit their results? It is possible that they are using contemporary natural language for their interrogation, and after not obtaining successful results, they then revert to the internet where contemporary natural language is prominent.

The online thesaurus of contemporary natural language search terms question produced interesting results. The participants' opinion showed they were fully in favour of the idea, especially if it linked to databases and other resource materials. Eighty percent of participants are in favour, and 20% are partially in favour. Further research is needed to investigate the possible realization of the contemporary natural language thesaurus as it would be highly beneficial not only to domestic students, but also to international students by enhancing academic skills and knowledge of the English language which in turn will assist research methodology and lifelong learning.

Recommendations

There are three recommendations.

- The first recommendation is that this survey should be conducted at the University of Canberra Library again in twelve months time with additional questions pertaining to the use of the AARLIN portal and the interrogation methods that students employ in using databases and the internet.
- The second recommendation is that an investigation should be conducted on upgrading library catalogues, databases and web pages within the context of the viability of implementing semantic web technologies.

- The third recommendation is to conduct research on the educational impact of semantic web technology and investigate the possibility of developing a contemporary natural language thesaurus.

Programming that represents the semantics of documents within web applications will provide a means of intelligent research that will more closely parallel the 'natural language' research processes of humans.[16] Developing a contemporary natural language thesaurus that interrogates all databases, thus saving researchers having to use different interrogation methodologies in different databases, should promote efficiency and enhance search success.

Conclusion

This paper has reported the results of a University of Canberra Library survey of student research knowledge skills, tools and resources, and has also investigated and evaluated usage patterns and abilities and has made recommendations for further investigations and research. There is sufficient evidence in the survey of the University of Canberra Library and the literature that higher education students need to have digital research tools and methodologies made more consistent and complementary. A semantic contemporary natural language thesaurus could assist in this.

This research has suggested that students lack knowledge of terminologies and of research resources and use tools that they are comfortable with in order to obtain information. Although the Lindsay and McLaren[17] research came to the same conclusion, the survey results also showed that even though they value the internet as a search facility, they also felt it was too easy for the researcher to be side tracked. Therefore, with semantic technology used for academic research on the internet, researchers should obtain more focused results.

Finally, because the great majority of students interviewed in the University of Canberra Library survey supported the idea of an online contemporary natural language thesaurus, the implementation of a semantic contemporary natural language thesaurus linked to databases and other resource tools is worthy of further investigation.

Notes

1 *Oxford Dictionary* 2nd ed Oxford Clarendon Press

2 S Irvine 'Essential Information Skills in a Busy Library' *School Libraries in Canada* vol 22 no 4 2003 pp33-34

3 W Feitzer 'Integrating Metadata Frameworks into Library Description' in C F Thomas (ed) *Libraries, the Internet and Scholarship: Tools and Trends Converging* New York Marcel Dekker

4 W Lindsay and S McLaren 'The Internet: An Aid to Student Research or a Source of Frustration?' *Journal of Educational Media* 2000 vol 25 no 2 pp115-128

5 J Field 'The Adult Learner as Listener, Viewer and Cybersurfer' in J Field (ed) *Electronic Pathways: Adult Learning and the New Communication Technologies* Leicester NIACE

6 S D Knapp, L B Cohen, D R Juedes 'A Natural Language Thesaurus for the Humanities: The Need for a Database Search Aid' *Library Quarterly* vol 68 1998 pp406-430

7 *Ibid*

8 *Ibid*

9 *Harrods Librarians Glossary and Reference Handbook* 9th ed R Prytherch (comp) Aldershot Gower

10 *Ibid*

11 *Ibid*

12 *Ibid*

13 Lindsay and McLaren *op cit*

14 *Library of Congress Subject Headings* 26th ed Washington Library of Congress

15 *The Contemporary Thesaurus of Search Terms and Synonyms: A Guide For Natural Language Computer* Searching 2nd ed Phoenix Arizona Oryx Press 2000

16 F Cervone 'W3C Delivers Standards for the Semantic Web' *Information Today* www.infotoday.com/newsbreaks/nb040216-1.shtml [accessed 4 May 2004]

17 Lindsay and McLaren *op cit*

Wilson Web

COPYRIGHT INFORMATION

TITLE: Natural Language Thesaurus: A Survey of Student Research Skills and Research Tool Preferences
SOURCE: Aust Acad Res Libr 35 no2 Je 2004
WN: 0415704324005

The magazine publisher is the copyright holder of this article and it is reproduced with permission. Further reproduction of this article in violation of the copyright is prohibited.

Copyright 1982-2004 The H.W. Wilson Company. All rights reserved.

RESEARCH BRIEF

MULTILINGUAL THESAURI FOR THE MODERN WORLD – NO IDEAL SOLUTION?

KERSTIN JORNA *and* SYLVIE DAVIES
{k.jorna, s.davies}@rgu.ac.uk

School of Information and Media
The Robert Gordon University
Aberdeen AB10 7QE

In the 21st century, multilingual tools are gaining importance as increasingly diverse user groups from different cultural and linguistic backgrounds seek access to equally diverse pieces of information. The authors of this paper believe that most current forms of multilingual information access are inadequate for this role, and that a new form of multilingual thesaurus is required. The core of this paper introduces their pilot thesaurus InfoDEFT as a possible model for new online thesauri, which are semantically structured, encyclopedic and multilingual. The authors conclude that while the manual construction of such thesauri is labour intensive and hence costly, pilot thesauri can be used as training sets for artificial learning programmes, thus increasing their volume considerably at relatively little extra cost.

1. INTRODUCTION

This research brief discusses thesauri as tools for multilingual information retrieval and cross-cultural communication. In an unpublished earlier version of this paper, we argued that although we agree with Chachra that 'multilingual subject searching has no easy and obvious solution' [1], an ideal solution to the complex problem of multilingual thesaurus construction would be the establishment of a language-independent knowledge base (see also [2]), which would form a conceptual structure common to all access languages. While we still believe that the construction of such a knowledge base would be extremely useful, work on our trilingual pilot thesaurus InfoDEFT convinced us that it is too ambitious a goal for a small team of researchers. Furthermore, such a thesaurus would be useful for the retrieval and understanding of facts, such as represented in Schmitz-Esser's EXPO 2000 thesaurus [3], but would not draw attention to linguistic and cultural differences between access languages. Close inspection of some terms particular to the field of ILS made us realise how important these differences are for cross-cultural communication. We therefore decided on the alternative solution of having language- and culture-based facets, i.e. structures that differ for each language thus drawing attention to differences not just in vocabulary but in conceptual contexts.

Journal of Documentation, vol. 57, no. 2, March 2001, pp. 284–295

In this paper we shall justify our choice of structure and layout by briefly presenting the complexity of the problem our thesaurus addresses, and by then describing the solutions offered by Schmitz-Esser and by ourselves. A worked example in Section 4 illustrates how InfoDEFT treats the differences between the apparently equivalent English term 'documentalist', French term 'documentaliste' and German term 'Dokumentar'.

2. ESSENTIAL THESAURUS FEATURES FOR MODERN APPLICATIONS

Three aspects are essential for good modern thesauri: firstly, they should be multilingual to support access to information resources not published in one's native language, and in order to facilitate cross-cultural communication in an increasingly global information society [4–6]. Secondly, they should be semantically structured rather than alphabetically [7]. This is partly to allow equal representation of all languages (rather than using one language as filing language, see also [6, 8]), and partly because semantic structures make the conceptual context of each term and its translations more explicit than randomly sorted lists. Thirdly, they should assist user comprehension and hence recall and precision in information retrieval by offering definitions, scope notes and even short encyclopedic articles. We shall discuss the first two points in more detail.

2.1. The need for multilingual thesauri

While modern information and communication technology facilitates global data transfer, it does not thereby facilitate global communication. Currently, both the Internet and international businesses largely use English for international communication, which poses problems of cross-cultural communication for both native and non-native speakers of English [4, 5]. Furthermore, English may soon cease to dominate global communication. Instead, language groups from Asia Pacific and Latin America are predicted to dominate the Internet [4]. New multilingual thesauri will be required to support communication between these language groups as well as between them and currently dominating Western European languages. Unfortunately, many so-called multilingual thesauri for subject access on the Internet or in library OPACs are in fact monolingual thesauri in conjunction with bilingual or multilingual dictionaries. These forms of multilingual information retrieval systems frequently fail, as can be demonstrated by a close inspection of *LISA* (*Library and Information Science Abstracts*). The *LISA* thesaurus contains only a minority of foreign entry terms although it contains abstracts and full texts from more than sixty-eight countries. Because all abstracts are translated into English, this leads to grossly inconsistent search results when there are no exact equivalents across languages. The French entry 'documentalistes', the German entry 'Dokumentar' and the English entry 'documentalists' for instance are translated as equivalents but lead to different search results (55, 4 and 110 records respectively). Following *LISA*'s thesaurus, one can use the term 'information professionals' as the preferred term for 'documentalists', which leads to 559 records. Using the term 'information professionals' without consulting the thesaurus, however, leads to 3,372 records.

Inconsistencies such as these can be overcome by offering equal access to all languages rather than using English as filing and main access language. This not only means equal treatment of cultures and hence of people [6, p. 84], but also allows unbiased treatment of the subject matter of the thesaurus, rather than of just one language's and hence one culture's treatment of that subject area [6, 8]. Unfortunately, it is not easy to achieve equal treatment of languages. There used to be physical constraints to the simultaneous display of several languages, which can now be overcome in electronic publications [6, 9]. Technology cannot overcome the intellectual problem of semantic equivalence, however. The complexity of natural languages, which all terminology is part of, rarely allows for one-to-one equivalences between terms of different languages. This is of little consequence to alphabetical thesauri, provided they allow for multiple translations and lexical gaps. It is, however, highly problematic for conceptually structured thesauri. Nevertheless, the following section argues that a good thesaurus should not be ordered alphabetically.

2.2. *The importance of explicit conceptual structures*

By offering facets or other semantic structures, a thesaurus aids users in their understanding of how terms are related to each other by underlying conceptual relations. This has two purposes. Firstly, it improves the understanding of the meaning of a term within its subject area, and can thus be used for didactic purposes (introducing students to the subject matter of the thesaurus, for instance). Secondly, it can enhance query formulation in information retrieval systems. Because users find it easier to recognise and select query terms or concepts than to formulate them themselves [10], explicit query expansion such as offered by a semantically structured thesaurus can be successfully used to trigger the memory of searchers looking for particular items (thus making use of the semantic priming effect).

For a multilingual thesaurus there are three ways of implementing facets or other semantic structures. Firstly, one can create a conceptual structure that is common to all languages to be included, an option recommended by international guidelines [8] and Hudon [6] but too ambitious for most applications unless an existing classification scheme is adopted (and most classification schemes will themselves stem from a particular cultural background). Alternatively, one can adopt the structure suggested by one of the languages in the thesaurus for all languages, which is the more practical approach adopted by Schmitz-Esser [3], but which favours one language and hence culture over the others [6, 8]. For our own thesaurus InfoDEFT, we chose a third way whereby we retain the conceptual structures suggested by the use of terminology in each language within five broad categories. The advantage of this approach is that linguistic and cultural differences can be fully displayed. 'For only faceted thesauri provide the flexibility of simultaneously presenting several possibly contradictory viewpoints' [11, p. 184].

3. TWO IMPLEMENTATIONS OF THESE ESSENTIAL FEATURES

We shall now describe and compare two thesauri. The multilingual thesaurus developed by Schmitz-Esser and his team for the EXPO 2000 in Hannover [3] is

a general knowledge thesaurus containing five languages. It overcomes the main difficulties of most previous multilingual thesauri. Our own pilot thesaurus is trilingual and restricted to ILS terminology. Both thesauri are electronic.

The main difference between these thesauri and previous thesauri is that they both go beyond the basic semantic relations of broader, narrower and related terms. Schmitz-Esser offers thirteen subdivisions of hierarchic relationships [3, p.14], and our thesaurus is faceted. Replacing the vague associative relationship of 'related terms', Schmitz-Esser introduces a 'Basic Semantic Reference Structure' (BSRS) containing six categories (CONCEPT, NAME, EVENT, LOCATION, ASPECT, TIME). These partly overlap with our own main categories of THINGS, PEOPLE, ACTIVITIES, PLACES, ENVIRONMENT. Both thesauri therefore show in which sense terms are semantically related. We agree with Schmitz-Esser's categorisation for a general thesaurus, even if we prefer our own approach for the narrower context of an ILS thesaurus. We disagree with Schmitz-Esser's, however, when he proposes to fill lexical gaps with paraphrases, where necessary [3, p. 12]. Contrary to Schmitz-Esser, we believe that the example of 'documentalist' shows that lexical gaps ought to be made explicit rather than hidden. Raising users' awareness of substructures that exist only in some languages enhances their conceptual understanding and should therefore improve precision in information retrieval tasks and communication.

Our approach differs from that of Schmitz-Esser in a second and more essential respect. While he enters all descriptors into a two-dimensional grid-structure of thirteen types of hierarchical structures in the one dimension and six elements of the BSRS in the other, our thesaurus is genuinely faceted in the sense that subfacets will be particular to their main facets. Our thesaurus thus follows a tree format (including many horizontal connections between 'branches' of different facets), while Schmitz-Esser's is organised in a table. His thesaurus layout may have the advantage of greater clarity and consistency, especially as users become familiar with the basic categories (BSRS). On the other hand, our thesaurus follows an internally related and, we hope, intuitive approach.

In conclusion to this comparison, Schmitz-Esser's thesaurus offers an encyclopedic overview and multilingual retrieval of facts, possibly at the expense of linguistic accuracy, while our thesaurus is aimed at enhancing the understanding of the precise meaning of terms in different languages, possibly at the expense of a general encyclopedic overview. The different functions of these two thesauri justify, we believe, these different approaches.

4. WORKED EXAMPLE

We shall now describe our pilot thesaurus InfoDEFT (Information Deutsch-English-Français Thesaurus) in more detail, referring to five sample pages as illustration (see pp. 288–292). We should like to stress that this is still a pilot thesaurus, which will be continuously expanded, altered and improved upon, following responses from users as well necessary changes due to the larger and more detailed vocabulary to be included. For easy expansion, alteration and maintenance all entries are fields in an Access database. This is accessible to users through an HTML interface (a webpage) thus allowing easy navigation via familiar underlined keywords.

Home	Things	**People**	Activities	Places	Environment	Deutsche Homepage	Accueil

People most relevant to the concept of information

1. information professionals
handling documents
processing information
delivering services to clients
applying ICT

2. authors and users of information in
education and research
tion
knowledge industries
media
communities
governments

Mensch
personnes

Information professionals processing information:

documentalists
information analysts
strategic information officers
knowledge managers
metadata managers

These are also referred to as:
information specialists

information scientists
information development officers
information managers
information officers

Dokumentar
documentaliste

Things: information, documents

People: librarians, archivists

Activities: indexing, abstracting, classification, documentation

Places: resources centres, informa-

centres

Society: professional associations

In the UK, there are no clear divisions between the professions named in this section, as job titles constantly change. However all of the professionals listed in this section handle information contained in documents rather than the documents themselves (a task reserved for the broad profession of librarians, as well as archivists and related professions).
In GB werden die Informationsspezialisten nach Tätigkeitsarten eingeteilt, und enge Berufsdefinitionen sind selten. Auf dieser Seite werden die wichtigsten Berufe genannt, welche mit Informationsinhalten umgehen.
Au Royaume Uni, il ne semble pas qu'il y ait de distinctions nettes entre les professionnels cités ci-dessus tant qu'ils sont compétents dans la pratique de la documentation.

Information professionals dealing with information in documents: documentalists – documentaliste – Dokumentar. Worked example page 1

Home	Things	**People**	Activities	Places	Environment	Deutsche Homepage	Accueil

Information professionals processing information are

documentalists
information analysts
strategic information officers
knowledge managers
metadata managers

These are also referred to as:
information specialists
information scientists
information development officers
information managers
information officers

Dokumentar
documentaliste

Documentalists *can also be referred to as*

Information officers
or by the general term
information professionals

Dokumentar
documentaliste

Things: information, documents

Activities: indexing, abstracting
classification
Places: libraries, resources centres,
information centres

Society: professional associations

"One who practices documentation. An information officer or intelligence officer who is concerned with the collection and dissemination of knowledge, rather than the librarian who is concerned with the techniques of handling records of knowledge [...] One concerned with assembling information contained within documents together with data from other sources to form a new compilation" [12]. For more information click here [13].
Menschen, die Wissen und Informationen sammeln und verbreiten, häufig indem sie diese mit Daten aus anderen Quellen als neue Sammlungen herausgeben.
Un professionnel compétent dans la pratique de la documentation (identification des sources de l'information, représentation, mise en mémoire et recherche de l'information).

Worked example page 2

Homepage	Ding	**Mensch**	Tätigkeit	Ort	Umfeld	English Homepage	Accueil

Informationsspezialisten

1. Der engere Informations- und Dokumentationsbereich
Archivar
Bibliothekar
Dokumentar

2. Verwandte Berufe:
Informatiker
Informationsvermittler
Informationswissenschaftler
Kommunikation- und Medienspezialist
andere Berufe
3. Benutzer und Verfasser von Information
Buchhandel- und Verlagswesen
Massenkommunikation
Allgemeine Öffentlichkeit

People
Personnes

Dokumentar

Unterschieden nach Ausbildung
Diplom-Dokumentar
Dokumentationsassistent
Wissenschaftlicher Dokumentar

Unterschieden nach Tätigkeitsbereich
Chemischer Dokumentar
Mediendokumentar
Medizinischer Dokumentar

information professionals processing information
documentaliste

Ding: Dokument, Informationssystem

Tätigkeit: Erfassung, Erschließung, Verwaltung, Dienstleistungen

Ort: Archiv, Bibliothek, Fachinformationszentrum,

Umfeld: Ausbildung
Informations- und Medienwirtschaft
Massenmedien
Wissenschaft
Vereine und Verbände

Dokumentare bewältigen verschiedene Aufgaben, die "alle mittelbar oder unmittelbar dem Informationstransfer dienen, wobei Information im dokumentarischen Sinne dasjenige faktische, intersubjektive Wissen ist, welches in einer aktuellen Problemsituation zu deren Bewältigung benötigt wird". [14]
Professionals who serve the transfer of information, where information is understood as factual intersubjective knowledge for the solving of specific problem situations.
Professionnels engagés dans la diffusion de l'information, dans la mesure où celle-ci est comprise comme étant une représentation des savoirs qui contribuent à résoudre des situations problématiques.

Worked example page 3

Accueil	choses	**personnes**	activités	espaces	environnement	Deutsche Homepage	English Homepage

Les professionnels de l'information

1. dans la gestion de documents
travail technique
gestion des supports
communications avec les publics

2. dans la gestion de l'information contenue dans les documents
travail technique
gestion des supports
de
communications avec les publics

3. dans la création et l'utilisation de l'information
education/recherche
l'édition
les médias
les communautés
les organes governementaux

Mensch
People

*Les professionnels engagés dans le **travail technique** de l'information incluent:*

documentaliste
analyste/indexeur
gestionnaire de language documentaire
bibliographe
autres professionnels

Dokumentar
information professionals
processing information

choses: information, documents, supports

activités: chaine documentaire, veille

espaces: *centres de documentation*, bibliothèques

environnement: formation, carrière, organisations professionnelles, sciences

la documentation

Ce groupe de professionnels engagés dans **le travail technique** de l'information possèdent le compétences spécifiques à l'information et la documentation, des savoirs et savoir-faire que tout professionnel du secteur doit posséder (identification et validation des sources d'information, représentation de l'information, mise en mémoire de l'information, recherche de l'information etc.). [15]
Professionals involved specifically with the processing of information contained in documents.
Informationsspezialisten, die sich speziell mit der Verarbeitung von Information in Dokumenten beschäftigen.

Worked example page 4

Accueil	choses	**personnes**	activités	espaces	environnement	Deutsche Homepage	English Homepage

Les professionnels engagés dans le travail technique de l'information (contenue dans les documents) incluent:
documentaliste
analyste/indexeur
gestionnaire de language documentaire
bibliographe
autres professionels

Dokumentar

Différents types de ***documentalistes***
documentaliste formateur
documentaliste de l'audio-visuel
documentaliste spécialisé
ingénieur documentaliste

emplois faisant appel partiellement aux mêmes compétences
bibliothécaire-documentaliste
correspondant documentaire
médiathécaire

Dokumentar
information professionals processing information

choses: information, documents, supports

activités: chaine documentaire, veille

espaces: centres de documentation, bibliothèques

environnement: enseignement, formation, carrière, statut, organisations professionnelles

Un documentaliste met à la disposition des demandeurs d'information ou des utilisateurs potentiels (sur demande ou sur sa propre initiative) les documents, extraits de documents ou données, conceptuelles ou factuelles, satisfaisant leurs besoins d'information, à un coût et dans des délais raisonnables. Assure pour cela la constitution et maintenance d'un fonds documentaire et des outils de recherche adaptés. [4]
Generic term describing information professionals who manage information and documents with the ultimate view to address information needs of potential clients. This may involve some skills which in the UK are specific to the work of librarians.
Allgemeine Bezeichnung für Dokumentare und andere Informationsspezialisten, deren Arbeit am Ende für spezielle Benutzergruppen oder Einzelkunden zugänglich gemacht wird.

Worked example page 5

Layout: With the exception of the introductory pages, every page in InfoDEFT appears as follows: a top menu bar contains eight choices, three for the home pages in the three entry languages and five for the main categories. In English these are called THINGS, PEOPLE, ACTIVITIES, PLACES and ENVIRONMENT. For easy orientation, these category fields are coloured and, once selected, their colour becomes the border colour of the four spaces below. There are three columns, two containing the broad and narrow context of the search term within the relevant facet, and one providing links to related terms in other facets and categories. The fourth space is a wide box at the bottom of the screen providing definitions and notes on the narrow context of the search term as well as links to encyclopedic articles (where appropriate). The advantages of this tabular layout are as follows:

1. users can immediately see the whole context of the search term;
2. the display allows easy navigation in all directions without loss of the previous context: selecting a term in the left (broad) column leaves this column unaltered while filling the other spaces with terms, notes and relations relevant to the search term; selection of a term in the middle (narrow) column copies the original middle column into the left column while leading to new middle, right and bottom entries;
3. there are links to nearest equivalent pages in the other two languages, so that users can immediately compare translations of their search terms and their respective broad contexts; furthermore the definition and notes given in the bottom section are summarised in the respective other two languages, so that users with little knowledge of the search language can still understand the definition and hence the meaning of the search term.

Encyclopedic pages are text documents containing alternative definitions, longer explanations regarding the history and current use of key terms in different user communities (if the term is used differently in North America from in Britain, for instance), a brief bibliography and links to other web sites (professional and academic organisations, online articles etc.). Co-operation with publishers of relevant encyclopedic reference books would allow this to be a standard addition for all terms.

Keyword search (to be added): Although the thesaurus is mainly aimed at navigation, we decided to offer a keyword option for quick access. If an entry term is keyed in that is not a preferred term, the user will arrive at the same page as through navigation for the equivalent preferred term. Until our keyword search is developed, users will have to begin new searches by new navigation from the menu bar.

Content: Documentalists, documentalistes and Dokumentare. The obvious difference between the entries in the three languages should not be dismissed as lack of consistency. By working from monolingual sources, all three entries were arrived at independently from each other and hence reflect what can be found in current literature. The advantage is that even without reading the note at the bottom of each screen users can see the difference between the three terms. We also decided to keep the different structures to see whether users clearly prefer one kind of structure to another (arriving straight at the names for the main

professions, as in the German pages, or navigating the tasks as in the French and English pages, for instance). If users show no clear preference, the differences in structures will be retained, as they are the most interesting aspect of the thesaurus for researchers. It is these differences that lead to focused research questions regarding the nature of ILS in different countries. Possible causes or at least patterns of differences may be arrived at by comparing the professions to their related activities. InfoDEFT could thus be the initial step towards answering questions such as whether the meaning of 'documentation' has changed in the same way as the meaning of 'documentalist' has, and whether these detailed observations are genuine insights into the nature of ILS as a whole.

5. CONCLUSION

We conclude with the optimistic outlook that there are practical solutions to multilingual subject access. Schmitz-Esser's thesaurus for the EXPO 2000 and, on a much more modest scale, our own thesaurus InfoDEFT are just two examples of new multilingual thesauri for the 21st century. Co-operation between the ILS community as classification specialists and the AI community as language processing specialists should result in a new format of macrothesauri that could include specialist and general thesauri and that would genuinely support international communication and knowledge transfer.

REFERENCES

1. Chachra, V. Subject access in an automated multithesaurus and multilingual environment. In: *Automated systems for access to multilingual and multiscript library material. Proceedings of the Second IFLA Satellite Meeting, Madrid, August 18–19 1993*.
2. Von Hahn, W. and Angelova, G. Combining terminology, lexical semantics and knowledge representation in Machine Aided Translation. In: *TKE'96, Terminology and Knowledge Engineering*. Frankfurt: INDEKS-Verlag, 1996, 304–331.
3. Schmitz-Esser, W. Thesaurus and beyond: an advanced formula for linguistic engineering and information retrieval. *Knowledge Organization*, 26(1), 1999, 10–22.
4. Cyberatlas. *The language of the web*. http://cyberatlas.internet.com/big_picture/demographics/article/0,1323,5901_150171,00.html (visited 5 October 2000).
5. Graddol, D. Will English be enough? In: Moys, A., ed. *Where are we going with languages?* London: Nuffield Foundation, 1998, 24–33.
6. Hudon, M. Multilingual thesaurus construction: integrating the views of different cultures in one gateway to knowledge and concepts. *Knowledge Organization*, 24(2), 1997, 84–91.
7. Riggs, F.W. Onomantics and terminology. Part I: their contributions to knowledge organization. *Knowledge Organization*, 23(1), 1996, 25–33.
8. International Organization for Standardization (ISO). *Documentation – guidelines for the establishment and development of multilingual thesauri*. ISO 5964:1995. Geneva: ISO, 1995.

9. Cabré, M.T. et al. A hypermedia project on terminology. In: *TKE'96, Terminology and Knowledge Engineering*. Frankfurt: INDEKS-Verlag, 1996, 424–430.
10. Iivonen, M. and Kivimäki, K. Common entities and missing properties: similarities and differences in the indexing of concepts. *Knowledge Organization*, 25(3), 1998, 90–102.
11. Priss, U. A graphical interface for conceptually navigated faceted thesauri. In: *Structures and relations in knowledge organisation. Proceedings 5th International ISKO-Conference, Lille, 25-29 August 1998*, 184–190.

Sources and notes for the worked examples: (unless otherwise indicated, definitions and translations are our own)

12. Prytherch, R. *Harrod's librarian's glossary*. 8th ed. Aldershot: Gower, 1995.
13. Clicking on 'here' will lead to an encyclopedic page citing Feather and Sturges on the term 'documentation': 'An approximate synonym for special library work or information science. It is the preferred term in the French tradition of information work. The word is not often used in English with the same broad significance that it has in French, although the prestigious *Journal of Documentation* is an exception to that general rule.[...] A full conspectus of documentation from a French publication, such as that in the *Encyclopaedia universalis* (1976), reveals the essential identity of the subject with the anglophone equivalents. The Special Library Association's Documentation Division was founded with a rather narrower definition, which claimed that 'Documentation is the art comprising of: (a) document reproduction; (b) document distribution; and (c) document utilisation'. (Feather, J. and Sturges, P., eds. *International encyclopedia of information and library science*. London: Routledge, 1997.)
14. Deutsche Gesellschaft für Informationswissenschaft und Informationspraxis e.V. (DGI). http://www.dgd.de/ (visited 2nd October 2000). A link to this page will be provided in the online InfoDEFT.
15. Association des professionnels de l'information et de la documentation. http://www.adbs.fr/ (visited 2nd October 2000). A link to this page will be provided in the online InfoDEFT.

Worked example: other examples and an update pilot thesaurus are available at: http://www.rgu.ac.uk/~sim/infodeft/ehomepage.htm

(*Revised version received 11 October 2000*)

A Metastructure for Thesauri in Archeology

Martin Doerr[1] and Demetrios Kalomoirakis[2]

[1]Institute of Computer Science, Foundation for Research and Technology – Hellas,
Heraklion, Crete, Greece
martin@ics.forth.gr
[2]Greek Ministry of Culture, Athens, Greece
Demetrios.Kalomoirakis@damd.culture.gr

Abstract. Controlled terminology becomes increasingly important for archeological documentation and information retrieval in wide area networks. Even though there is a vast literature on thesaurus creation, the logical organization, the "metastructure" of a thesaurus is still rather unreflected, in particular in the more abstract layers. So-called facet indicators are distributed intuitively in order to group a set of sibling terms by an implicit property. They can be regarded as potential elements of a metastructure. The Greek Ministry of Culture is developing a general purpose thesaurus for archeology in cooperation with ICS-FORTH. Based on contextual and phenomenal notions, an extensible, systematic metastructure for terminology in archeology is proposed, which seems to be more comprehensive than traditional approaches. This paper discusses the benefits of such an approach. In particular, the proposed structure seems to provide a better understanding of the implicit criteria intrinsic to object classification terms and the relation between compound expressions and traditional terms.

Keywords: archeological terminology, facetted classification, thesauri, qualia, subsumption hierarchies, concept formation

1 Introduction

Controlled terminology becomes more and more important for the electronic documentation of archeological object description and its consistent retrieval in wide area networks. In particular thesauri, which organize terms or concepts in semantic networks, become a mandatory tool for users to find their way through the rapidly growing electronic information flood and for unambiguous communication through automated intermediaries. Even though there is a vast literature and established practice about creating thesauri (see e.g. publications of the Getty Foundation, the British Arts and Humanities Data Service AHDS, and numerous national standards), their logical organization, the metastructure of a thesaurus is still rather unreflected, in particular in the more abstract layers. In this paper, we try to clarify some concepts and present the logical organization of a general purpose thesaurus for archeology developed at the Greek Ministry of Culture in cooperation with ICS-FORTH. Finally we point out some research topics, that came up with this work.

2 What is a Thesaurus

The notion of a linguistic thesaurus has been introduced more than a hundred years ago by Peter Mark Roget in 1852, with his famous "Thesaurus of English Words and Phrases", for dictionaries intended to help authors find synonyms or more precise terms they have not in mind. Such thesauri are ever since quite popular, and Roget's one is still being edited [Encarta].

With the raise of electronic communication, and in particular the search of information in database records, the need to refer to some concepts with standardized expressions also arose, else entries made by curators and requests done by users would never match. The currently most successful approach to standardize expressions are so-called thesauri in the sense of computer science.

Two directions can be distinguished at the moment: A linguistically oriented and a conceptually oriented. The "linguistic" one regards the "term" as the key element of the thesaurus, coming from a dictionary world. Terms are represented as the expert would use it in speech: potsherd, vessel, column, pottery etc. Besides textual descriptions of the meaning of the term, linguistic information is added, like part-of-speech, sort keys etc..

Respective data standards have been promoted by the Text Encoding Initiative TEI [TEI], the latest developments being ISO 12620, MAchine Readable Terminology Interchange Format MARTIF, and the Virtual HyperGlossary VHG (Murray 1998 : 34). Terms are typically organized in monohierarchies (i.e. one term belongs exactly to one broader term) as a kind of associative decision trees leading from higher to lower

abstraction levels of conceptually close terms. These hierarchies serve mainly as searching aids for the user to find a term rather than expressing abstraction levels. Often information is added to create a nice lay-out of indented lists in printed reports.

These thesauri might be helpful for individuals looking up and understanding terminology as an encyclopedia, for linguistic work, translations and natural language search. For indexing and classification in databases however there are severe drawbacks. The naïve assumption that a term uttered by an expert is a sufficient surrogate for the item we are describing or looking for does not hold for the following reasons: Terms are context dependent, be it true homonymity ("contrastive ambiguity" (Pustejovsky, 1995 : 27), shift of meaning ("complementary polysemy"), or because the term expresses not all aspects of an item relevant for the one or other future user. A "vessel" can be quite different things, "pink" is not always a color etc. A "temple" of the Greco-Roman period may have different associations and abstractions from a Protestant one. "Neoclassical building" and "school house" can be equally relevant characteristics of the same object. Further, users often query at a higher level of abstraction than the most detailed classification of an object, e.g. "black vases from *Attica* of the 3rd and 4th century BC". Hierarchies that are not simply associative, but that are built on true abstraction, can help the software to resolve such queries automatically through the thesaurus.

Such considerations lead the library scientists in the middle of the 20th century to the development of classification languages based on concepts rather than terms. As C.Welty (Welty 1999 : 155) points out, only since the shelving of books on unique places is no more the major organisation principle, assignment of multiple subjects/terms were considered. So on one side, the problem was addressed by rules to combine "elemental" concepts either in the terminological system already (pre-coordination) or at classification time (post-coordination). The classification vocabulary was grouped for that purpose into so-called "facets", the sets of elemental concepts, going back to Ranganathan's Colon Classification system (Ranganathan, 1965). On the other side, these "facets" were internally organized by conceptual relations. Such systems form **thesauri** in the sense we will use from here on. As Pollitt (Pollitt a, 1997) refers it:

"Problems in the Verbal Plane brought about the introduction of regularisation and structure into the language used in post-coordinated systems via thesauri. Vickery (Vickery, 1960 : 181) reports how he first encountered the word thesaurus: "Speaking at the Dorking conference on classification, Helen Brownson said 'The problem [of information retrieval], as some investigators see it at least, is to transform concepts and their relationships, as expressed in the language of documents, into a somewhat more regularized language, with synonyms controlled and syntactic structures simplified. Now it is reasonable to think that the further we can go in routinizing and mechanizing the techniques of translating ordinary language into a regularized language and of coding for machine manipulation, the more we will be likely to achieve economically feasible machine searching on a large scale [Some investigators] have come up with the thought that the best answer may be the application of a mechanized thesaurus based on networks of related meanings.'" In the same paper Vickery (1960 p.185) cites Bernier and Heumann (1957) as proposing the introduction of an organised vocabulary in the form of a thesaurus".

This "conceptual direction" of thesauri, typically elaborating on the semantic links as defined in ISO2788, ISO2709 and ISO5964, regards the "concept" as key element of the thesaurus (see e.g. (Svenonius, 1989 : 82)). These concepts can be alluded to by a series of linguistic expressions, but the concept itself has no name. For better orientation of the user, one term may be selected arbitrarily as **identifier** for each concept (often called "**descriptor**"), giving eventually the impression as if the thesaurus were still about the term. This idea is also well explained in the explanations to the Art and Architecture Thesaurus AAT [Getty94], whereas ISO2788 still has no clear notion of separating between terms and concepts. WordNet (Miller 1993) identifies the concept (here called "meaning") by the combination of **all** appropriate terms. Miller et al. make clear the many-to-many relation between words or terms and concepts.

Finally it should be mentioned, that there is the automatic generation of so-called "thesauri" (e.g. (Hazewinkel, 1996), (Chen, 1993 : 25). By statistical means, measures for a conceptual distance between terms or expressions are derived based on the co-occurrence in corpora and by grammatical criteria. Those methods do not give much insight into the semantic nature of the derived relations, and should more be regarded as retrieval aids than thesauri. Nevertheless, in the future such methods may quite well assist in the creation of true thesauri.

3 About concepts

Cognitive psychology scientists have proposed several definitions for the notion of "concept" (see also [Doerr 1998). According to a first point of view, a concept is a person's conceptualization of the notion of categories

(the class or set of entities which are grouped together on the basis of some criterion or rule (Michalski, 1993). According to this point of view, there would have to be as many concepts for the same category as the different cognitive representations for it. For thesauri we do not adopt this approach since the thesauri incorporated are collections of terms with generally agreed on semantics and not individual definitions or representations of a concept.

According to a second point of view, a concept is perceived as a set of entities, called "concept instances" by common agreement rather than formal reasoning on the properties which may make an entity instance of the concept. In other words, we regard a concept as defined within a group of people, as long as the people agree which items are characterized by the concept. This does not exclude the use of rules, but it is not based on it. The difference to the first point of view is, that we do not argue about what people really conceptualize, but what they can agree on to the outside. This is actually the basis for semantic thesaurus construction. It is further a principle, which can be easily applied to the classification of material objects. Eventually we are forced to communicate such concepts to each other by a verbose account of examples and contextual information, called "scope note" in the context of thesauri.

According to a third point of view, a concept is identified by a classification rule. It means, that the concept is further analyzed into a logical expression about its properties. However quite often conscious rules cannot easily be formulated for even obvious concepts, as e.g. the optical characteristics of an aquarelle, and the rules themselves cannot be further analyzed. Nevertheless, this principle is the basis for the so-called "Terminological Logic" or "Description Logic"(Borgida, 1995 : 671), [DL], which can be regarded as the most advanced extension of conceptual thesauri. It has been proven quite successful in medical applications (Rector 1997) and other domains. To our opinion it is more appropriate for cases where distinct overt features dominate, like in scientific objects and technical artifacts. It is quite useful to complement an agreement-based body of concepts by the more dynamic rule-based expressions, as I will show below.

Advantages of the conceptual approach are, that conflicts of homonymity, and cultural and contextual differences can be overcome or made explicit. What counts is the definition of the concept rather than the term itself. Consequently, the AAT team has created expressions like "pink (color)", which are not natural, but resolve ambiguity. Further, we can identify clear abstraction hierarchies based on semantic inclusion (isA, generalization, superclass, subsumption, hyponomy and whatever synonym term may be in use), that allow to expand more abstract database queries into sets of specific terms. E.g. "dime" is a "coin", "baby doll" is a "recreational artifact", "kachina" is a "figural work" etc.

3.1 About Objects and Subjects

As we can see from the historical considerations above, library science, and in particular the problem of subject classification, has been the driving force behind these developments. Museums and archeology, dealing to a very large extent with material objects, have relatively unreflectedly adopted the same method to classify their objects instead of the literature about them. Let us make here some propositions about the differences:

- Physical objects are not "abstract". An "Introduction to Biology" and "The Behaviour of Geese" can be the distinct subjects of two different books. In case of objects, there is no one a "living being" e.g., without being something concrete, like "goose".
- Consequently, a physical object is truly classified by any higher abstraction of its type, but only the lowest has instances. My dog is a dog, a carnivore, an animal, a living being, a material object. Any existing animal however must be of the most specific type, whereas a book about animals may be as generic as the concept "animal".
- The lowest abstraction level is relative, a function of our attention to more or less characteristics. It depends on the size and constitution of the collection at our disposal. The relatively well-defined level of a "species" in biology has only weak analogies in archeological objects.
- An object may be classified by an open number of relevant views, according to the investigators objectives. These views may be completely uncorrelated as e.g. "exchange item", "caouri shell", and "property of an Oba", "bead", "element of a preserved string", "slightly worn". "Subject" on the other side may be seen as one such view, valid for literature objects, but mostly absent in material objects.
- Even though any object can give raise to a subject in the sense of "talking about " (Welty 1999 : 155), the "parallel hierarchy" (Soergel, 1995 : 369) generated by objects viewed as subjects may differ. E.g. talking about "bridge construction" is a kind of talking about "bridges". "Bridge construction" itself is not a kind of "bridges", but a kind of construction, either a design or an activity or both.

Of course, any piece of literature as a whole has its own unique individuality similar to a museum object (Welty b, 1999). We however agree here with (Welty 1999 : 155), who prefers to see the subject of a certain book not as a parallel individual associated with each book, but as a collective concept applicable to more than one book. He further points out, that in libraries fiction itself is mostly not classified with subjects. The focus of libraries and museums is obviously quite different, and give raise to endless misunderstandings between both communities.

Summarizing, we have the impression, that conceptual hierarchies for material objects can be and need be stricter defined with respect to the notion of abstraction (hyponomy, broader/narrower terms) than usual in library subject catalogues, and that they need more structure reflecting different aspects of investigation. E.g. we cannot declare, "Top: Arts: Classical Studies: Journals" and "Top: Arts: Classical Studies: Academic Departments" as narrower terms of "Top: Arts: Classical Studies" (see the dmoz project,[DMOZ]), or regard bridge construction as specialization of bridges (LCSH), but must separate entities of the real world by their deeply distinct nature.

3.2 Descriptive and Characteristic Concepts

The examples above suggest, that we may be able to separate concept hierarchies by the aspects they express, about people for instance by sex, race, intellectual properties etc. This can be done easily with concepts, which represent more or less verifiable properties of the one or the other kind, which **describe** basically an object by observational criteria imposed by the researcher. See e.g. the definition of "knives (weapons)" in the AAT: "Weapons designed primarily for cutting, consisting basically of a single-edged, pointed blade and a grip mounted asymmetrically in relation to the axis of the blade, closer to the back edge". The concept of "weapon" itself is purely functional, and bare of any assumption about its physical form. A dynamic (post-coordinated) term like "railway museum" expresses not more than what it says (see the AAT).

We have however simultaneously to do with historical concepts of things **characteristic** for a kind of "design model" of a culture, like an "ushabti" or "kachina doll". These concepts comprise a series of aspects and implications. "Persian rugs", e.g. exhibit the subtlety that not all rugs from Persia are Persian Rugs, nor are all Persian Rugs from Persia, even though at least a "genuine" one should be. Obviously the provenance "Persian" does not imply that a rug necessarily has the characteristic design observed in Persia, and the characteristic design may be copied or truly followed outside of Persia. The latter has a history, evolution and social context.

Let us in the sequence talk about **descriptive** object concepts, as those, that analyze our observations about the objects, that have come upon us from the past, be it on features that can be verified now on the item itself, or features that originate in historical knowledge.

In contrast, let us talk about **characteristic** object concepts, as those, that designate a "design model" of a culture, which has specific functions, designs and contexts of use altogether, has characteristic traditional names (e.g. "alcazar"), varies over regions and evolves over times. The "ushabty" e.g. was initially a surrogate for the mummy, and later became a servant for the afterlife. It has a characteristic span of forms and sizes. "Ushabty (ushebti, ushabti, shawabty etc.) are transcriptions from the original Egyptian terms for that concept. In other cases, like with neolithic objects, those are lost, and researchers "coin" modern terms as surrogates.

As we see, our object concepts come from different perspectives of cognition, with more complex relations between each other and the objects they classify, than simple hierarchies can appropriately capture. The results of pressing those in a simple hierarchy are disorientation of the user trying to find concepts and poor or even questionable conclusions from the given semantic relations.

Ranganathan (Ranganathan 1965 : 38), characteristically states: " that the work gets resolved into three different planes: the Idea, the Notional, and the Verbal Planes. Looking back from this position into the work done before these three planes of work were clearly seen and separated, it is found that much of the difficulty arose out of frequent, listless, and unconscious change from one plane to another. Another cause of difficulty was the inhibition of work in the Idea Plane by the limitations of the Notational Plane, and by the inherent defects of the Verbal Plane." (See also (Pollitt a 1997)).

In this paper, we want to investigate some properties of the Idea Plane, the world of concepts and meanings, and their relation to objects of our investigation, free of limitations of the Notational Plane, and in the sequence propose some improvement in the Notational Plane.

4 About facets

We have already mentioned the so-called "faceted classification". The term "facet" alludes to the fragmented surface of a diamond, suggesting that the same thing can be seen from different sides. Faceted classification and further developments in the form of Description Logic become more and more popular to deal with the complexity of precise capturing of information contents (e.g. Prieto-Diaz, 1987: 6, Constantopoulos 1995, Welty b, 1999 and Pollitt a, 1997). Actually one may distinguish three slightly different notions of facets with at least two different applications.

Following (Prieto-Diaz 1987 : 6): "The faceted method relies on building up or synthesizing from the subject statements of particular documents. By this method, subject statements are analyzed into their component elemental classes, and these classes are listed in the schedule. Their generic relationships are the only relationships displayed."

This is the first notion of facets: **Elements for synthesizing** complex (post-coordinated) terms or statements by enumeration. To our opinion, this kind of analysis results more in grammatical elements, like subject, predicate, object etc. (see e.g. (Constantopoulos 1995), than in "elemental classes". They become elements of an indexing language to create an open number of potential "compound terms". Typical library facets are: Topics, periods, places, genre.

Whereas statements often summarize quite well a scientific paper, material objects are better characterized by noun phrases. Soergel (Soergel 1995 : 369) discusses in much detail the use and limitations of an indexing language (the AAT), which foresees only one generic relation between each facet. An example for such a compound is "19th century Massachusetts wood chairs". Please note, that this denotes no more than a chair. The AAT distinguishes the facets: "Object, Agents, Activities, Styles and Periods, Materials, Physical Attributes, Associated Concepts". Description Logic is the natural extension of such indexing languages, as it enriches the set of relationships allowable between the concepts and puts them in a well-defined logical framework (see e.g (Welty b, 1999), (Bechhofer, 1999) about the utility of DL).

The separation of our concepts into **disjoint cognitive categories** gives raise to a slightly different interpretation of "facet". Under this aspect, we regard as facets notions of objects, actors, events, measures, time, space etc. "Facets" are seen as the building elements of our conceptualizations of a domain, without an immediate syntactic purpose. Such elemental classes are often referred to as "**major facets**". Ranganathan talks about "the five fundamental categories PMEST (Personality, Matter, Energy, Space, Time)", and the AAT facets seem to be of that kind. The difference to the above interpretation becomes obvious, when someone regards grammatical subjects and objects, which are identical in nature, but different in their role within a statement. E.g. a physical object can be subject and object in a grammatically correct phrase.

For an archeological thesaurus, we regard the principle of "major facets" as appropriate to bring a first order into all potential concepts of the domain. The ontology of the CIDOC Conceptual Reference Model (Doerr 1999) is a attempt to provide a formal standard model to describe objects and their history, mainly to be able to mediate and transform between different data and metadata formats. It distinguishes: **Temporal Entities** (including periods, events and activities), **Actors, Physical Objects, Conceptual Objects, Place, Time**. Recently the IFLA and the Dublin Core community come up with nearly identical categories (see e.g. the Indecs project, http://www.indecs.org/results/model.htm). The obvious similarity encourages standardization. Most probably behind are innate categories of human perception, as e.g. Steven Pinker (Pinker 1994) suggests. We would however not dare to say, to which degree nature itself dictates those. From these categories it is easier to derive facets for synthesizing complex terms for different applications, than from purely syntactically motivated categories.

There is however yet another interpretation of "facet", to our opinion completely different. (Prieto-Diaz 1987:6) continues in the same paper: "Facets are sometimes considered as perspectives, viewpoints, or dimensions of a particular domain", and continues with an example about animals classified by habitat versus genealogy. Actually this is the notion that Ranganathan originally introduced, which reminds indeed facets of a jewel. The MDA archeological thesaurus [MDA97] e.g. introduces a term "armour by construction", and below "scale armour" etc, "armour by form", and below "cuirass" etc., "armour by function" and below "parade armour". Such terms, that announce the criterion by which the subsequent analysis is carried out, are called node-labels (ISO2788), **minor facets** or guide-terms (AAT). Frequently, the values these criteria can take on are taken from another "major facet". As e.g. in the term "cutting sword", the value "cutting" appears in the major facet "Activities" of the AAT.

4.1 The Minor Facets in the AAT

Ranganathan observed, that an open number of concept arrays can be used to provide the criteria for a facet on another array. Similarly, the AAT has introduced guide terms (facet indicators) systematically in a great number, actually 2840 in the 1998 edition. From a scientific and philosophical point of view, this situation is not very satisfactory. In a first statistical investigation, we have identified, that there are quite well predominant criteria, that can be used to create a metastructure.

For that purpose, we have separated the guide terms of the object facet, about 1640 (some concepts are not really objects, as "illustrations", "museums"). From these, 615 actually refer the criterion with the characteristic connector "by" like "swords by function" (see Appendix 1). These could easily be separated, and became the object of our analysis. A more thorough analysis could well be done on the total, but we are here only interested in a qualitative result:

318 guide terms out of the 615 fall in three clearly distinct dominant groups (see Appendix 2):

by form :	129
by function:	121
by location or context:	68

The rest has frequencies below 15. The notion "location or context" is not easy to interpret. It is often a geometric relation of a part to a kind of whole, but also of independent smaller elements to a larger structure as *landscape* to sea, lamps to parts of a room etc. The other categories are clear cut. In practice, form and function can overlap in the form of "functional forms". There are 15 guide terms " by form or function", like "furniture by form or function" and others.

This analysis is only linguistic. Actually a closer look on the actual narrower terms appearing under the less frequent kinds of guide terms reveals more of the major categories, as " by shape", " by merchandising practice", " by position". The AAT has never tried to standardize these terms. Further, a large multitude of differently named categories can be characterized as construction criteria, be it the internal structure or the process of creation of the internal operation principles of machinery. We have not tried to go into more detail, nor have we tried to distinguish event-like functions like "cutting" from activity-like functions like "dwelling". Construction and form sometimes overlap, in the case of technically necessary forms of machinery we have preferred to classify the criterion as constructive.

We have arbitrarily added the group "form or function" to "function". There is a small group of criteria about the social context of creation or use. Surprising enough, only 5 times the "subject type" appears in the proper sense, e.g. for sculpture, as a term-generating criterion ! We have classified the "subject" of a museum e.g. under "function". Here the statistics of our interpretation of the AAT object guide terms with "by" connector:

Criteria of form:	34%
Criteria of function:	29%
Criteria of relative location:	15%
Criteria of construction:	15%
Criteria of social context	5%
Criteria of subject type	1%
Criteria of naming (like coins)	1%

From the impression these statistics give we derive the hypothesis, that there are a few predominant aspects to analyze objects, and not an arbitrary, unpredictable mass. These aspects are reflected in the terms themselves, but may be as well explicit criteria (distinct data base fields) of classification (function, form, material etc.). Some British museums have made the functional aspect explicit: they classify objects with SHIC, (Social, Historical and Industrial Classification) [SHIC93] only by their relevance to human activities - a clear indication of the relevance and utility of the "function" and "social context" aspect. In this practice, the actual "function" is always explicit, and not hidden in some traditional term.

The above analysis gets an unexpected confirmation from modern linguistics. Pustejovsky (Pustejovsky 1995) has developed a theory to cater for the dynamic change of meaning of words in new contexts. He shows, that nominals in general, and in all languages imply multiple meanings, which are "activated" in a sentence by the appropriate context, e.g. a verb that refers to a certain aspect. An example he gives is:

"He walks through the door" (function, opening), and "He paints the door" (object).

He talks about the "Qualia Structure" of nominals, which he analyses in the following categories (referring also to Aristotle's notion of modes of explanation):

- Constitutive: The relation between an object and its constituents. (Material, weight, parts and component elements)
- Formal: That which distinguishes the object within a larger domain. (Orientation, magnitude, shape, dimensionality, color, position)
- Telic: Purpose and Function of the object. (Purpose an agent has in performing an act. Built-in function or aim which specifies certain activities)
- Agentive: Factors involved in the origin or "bringing about" of an object. (Creator, artefact, natural kind, causal chain)

The similarity is striking. If finally common sense language understanding and classification in archeology lead to the same or only similar categories, further research will show. By sure, both disciplines can learn from each other.

5 A metamodel for object thesauri

The fact, that traditional terms silently imply the one or other aspect can raise confusion in classification. Some data about an object may be classified under the above aspects with descriptive concepts, especially if a characteristic term for it is missing. Another object may be classified with traditional characteristic terms, that imply some descriptive concept shared with the other object. We shall not be able to retrieve both under the same terms.

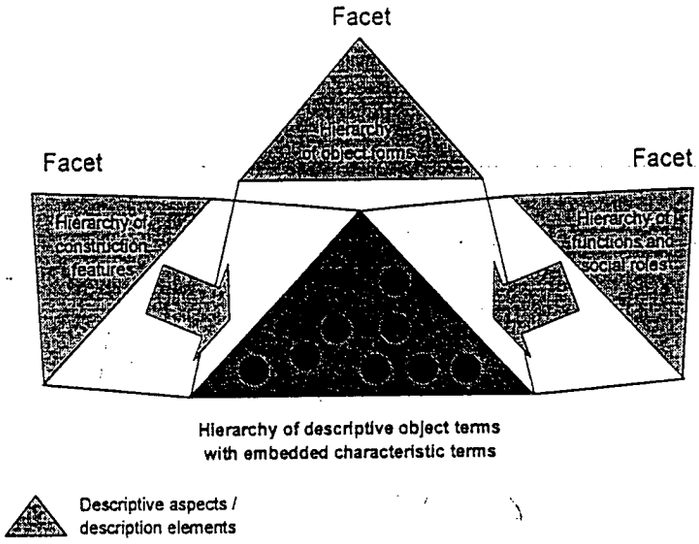


Figure 1 : Conceptual Scheme of a multi-faceted Hierarchy

See e.g. the AAT Terms "armor", "parade armor" and "garnitures". The scope note of "garnitures" says: "Sets of armor, for use in warfare or tournaments. They were invariably made for important persons according to a precise design for specific types of events." Neither the aspect of protection in combat nor the aspect of social representation are explicit from the wording or from the broader terms given. On the other side, a "pure descriptive" classification with the above criteria is neither efficient nor always appropriate. Archeologists at the Greek National Archive of Monuments and others [PrivCom] have repeatedly referred to us the problem, that the characteristic terms imply various properties not accessible in a database search. Explicit reference at each instance on the other side is not economic, and implicit properties may change with the period even for the same term.

Under this consideration we propose to classify systematically characteristic terms by descriptive terms. By this practice, direct classification of objects with descriptive terms and classification with the correct characteristic term can become consistent and compatible. As we have seen above, a thesaurus can serve as indexing language

and as an information source about a conceptualization of a domain. For the latter, systematic classification of descriptive terms by all aspects (qualia) is a remarkable finding aid for users, as detailed in chapter 6.

An object clearly identified as a model of its culture, like an ushebti, is of course best characterized by its vernacular, historical term. Only that will render all relevant aspects. An object, which cannot clearly be identified as the one or the other, should better be classified by all possible aspects, in order to provide future researchers with all potential candidates for their requests (ensure recall). Similarly, a term with multiple potential meanings should be generalized (i.e. assigned broader term relations) by all possible meanings. Equally prominent in archeology are cases, were we try to identify the unknown models of a culture, in particular of prehistoric ones. In that case, we may use combinations of descriptive terms and eventually invent a new characteristic term, a "coined term", when our knowledge about the kind of items seems to stabilize.

This is not in contradiction to defining its proper scientifically established meaning by respective scope notes. *This double* practice not only helps to explain vernacular terms or their equivalents to the interested scholar, but also provides the correct generalizations for query expansion in the use as indexing language. Even more, shifts in meaning could be expressed by specializing a term into its local and temporal variants, and generalizing them *differently*. E.g. most ancient Egyptian items come upon us are burial items, but their later equivalents are not [PrivCom], (contemporary secular equivalents are typically not preserved).

5.1 Examples

To illustrate the theory, let us in the following assume, that the relevant qualia are: Function, form, construction. To create the proposed metastructure, we may provide an initial vocabulary of expressions of function (like SHIC e.g.), of forms, and of constructive features relevant to the set of objects under consideration. From these, we create the set of classes of objects, which are characterized by precisely one feature of the initial vocabulary, e.g. "objects for warfare", "objects for eating & drinking" "rectangular objects", "built objects". If the initial vocabularies are organized in by hyponymy (broader/narrower terms), this procedure creates a "parallel hierarchy" (Soergel 1995: 369) for each aspect (quale). These would be the primitive descriptive concepts. Each hierarchy could be headed by a node label like "objects by function" etc., even though the practice is to enumerate only a flat list under each node label. Any characteristic object concept should be describable as narrower term of some combination of theses classes. In this theoretic picture, characteristic concepts appear dispersed in the "pool" of all possible combinations of the primitive descriptive concepts, as fig. 1 suggests.

In practice, two factors cause severe distortions to this picture: Properties of the formal aspect and the relations between the characteristic concepts. The formal aspect or quale seems to play a dominant role the discrimination of characteristic concepts. Pustejovsky refers it as "basic", without further justification. We observe a basic difficulty, to define expressive sets of shapes without reference to objects types, whereas we derive easily shapes from object types: we may talk about a cross-shaped ground plan, a hammer-like fish ("hammerhead shark") etc. This does not hold for the other qualia. This is further confirmed by the dominance of form criteria in the AAT. As described in chap. 6, the Greek National Archive of Monuments has introduced the formal aspect as «είδος», literally translated as "kind, model or article".

It is beyond the scope of this work to analyze the deeper reasons behind that. It seems be economic, to take characteristic concepts themselves per default as sufficient to represent their formal aspect. Consequently, an independent vocabulary of form is only created for those forms not captured by the characteristic terms themselves, and the total of characteristic terms is seen as a source to generate respective form expressions, e.g. with a "like" operator.

Beyond that we observe, that characteristic terms of some higher level of abstraction may be specialized themselves by one or the other aspect, or all aspects together. In the AAT, the guide terms appear characteristically at the top and then at any level without recognizable system. Our interpretation is that actually any characteristic term is potentially specialized by all aspects at any level, however these potential places in the polyhierarchical grid may not always be occupied with actual established terms. The current practice to introduce a node label in order to announce the aspect of making narrower terms (NT) has two limitations: First, it cannot be continued on the next level. Another node label is needed. Second, it does not say, which aspect the inversion, the broader term relation (BT), entails. This is actually non-trivial. There are combinations of all aspects between the two directions of a BT/NT relation. The broader term relation is actually the one, that classifies the characteristic concept, whereas the narrower term relation is the one we need to systematically create the thesaurus top-down and to expand higher concepts into individual classification concepts.

Instead of node labels, we may better introduce different types of BT/NT relations, a real metastructure feature, which goes beyond ISO2788. The idea is illustrated in fig.2, around the characteristic concepts "swords" and "foils (swords)". A type determinator in brackets accompanies the latter term, in order to disambiguate the concept from homonyms. The right branch (weapons,) shows part of an object hierarchy induced by activities "fighting and hunting", "cutting and thrusting". These pure functional concepts become specialized by the characteristic term "sword", which adds a morphological and constitutive aspect. The latter is again specialized by the functional aspect "fencing", and in turn by the characteristic concept "foils (swords)".

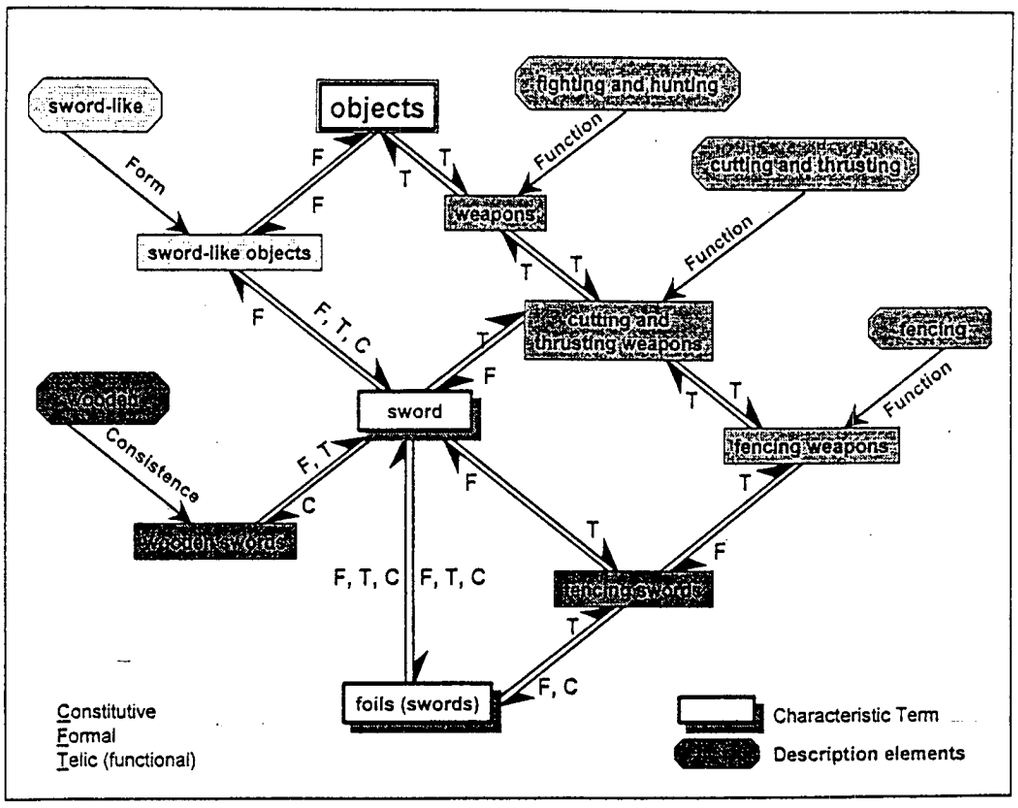


Figure 2 : Sample hierarchy with typed broader/narrower term relations

On the left side, we see the morphological aspect, starting from "sword-like", and penetrating through the concepts "swords" and "foils (swords)". "Fencing swords" takes a hybrid position. We show further a constitutive aspect, the "wooden swords". In the hierarchy here, wooden swords appear as weapons, and to our knowledge at least in medieval Japan, wooden swords were in use as respectable weapons. Alternatively, one could put wooden swords as a formal/constitutive hybrid under "sword-like objects", and distinguish between "wooden swords (toys)" and "wooden swords (weapons)". From this picture, we can easily recognize how the generalization or specialization has a direction towards the formal, the constitutive, and the functional aspect. A similar graph can be seen in (Pustejovsky 1995:145).

"Weapons" is an established, "historical" term and concept, which is purely functional. Obviously, we cannot assign any prototypical shape to "weapon", as we do to "sword". As such, we regard it as descriptive and not characteristic. However in general, non-formal terms are more frequently compounds than the others. In particular concepts which complement gaps in the lattice of traditional terms in order to cover relevant feature combinations are compounds with terms from the participating aspects, as expected from the faceted classification schemes presented above. The difference of our position here is that we cannot replace a part of the hierarchy by purely descriptive compounds as propagated by current indexing systems, but characteristic terms and traditional terms penetrate such hierarchies. A solution must be found, to embed those in a comprehensive way in a scheme of dynamic feature combination, as indicated in this paper.

6 Polydeykis, a Practical Experience

POLYDEUKIS is a general purpose, ontological Thesaurus aiming at classifying consistently cultural and archaeological terms. It is being elaborated by the Directorate of the Archive of Monuments of the Hellenic Ministry of Culture in cooperation with ICS-FORTH. It has reached its first state of completion, where the basic

principles can be verified and the further refinement can be determined. Its appellation was coined to honor Polydeukis, a second-century BC Alexandrian who authored a Lexicon of terms of the Attic dialect. The specific activities of the Directorate of Monuments have instigated a thematologically multi-faceted logical organization of the cultural and archeological terms contained in Polydeukis. Indeed, the Directorate handles an archive of extensive and multi-faceted data concerning cultural monuments. These are the material traces of coherent human activity spanning over several millenniums from a geographical area considerably larger than the Modern Greek State. The influences exerted by the Hellenic culture over other cultural environments with which it held converse and still does, are very broad and remarkably fertile.

Consequently, all objects under investigation are characterized by an exceptional variegation of terminological designation and distinction into characteristic concepts. The analytical overview of concepts in Polydeukis follows an anthropocentric logical structure, that is to say, it observes closely, albeit up to a certain degree, the relationship which man (as a living being) has with his fellow humans, and with the creations of himself and of nature. Every human being, for instance, is a biological entity designated by its parents (creators), the time and place of its birth, its name, its physiognomy and bodily form, its personality, and its biological structure, as well as its activities, and the ambient wherein it lives and works.

In analogy, the logical organization of Polydeukis establishes a conceptual anthropocentric structure for every type of object/monument, to which all characteristic historic terms designating and specifying it are linked through a system of guide terms. The analysis of characteristic concepts conducted so far yielded the formulation of a relatively repetitive conceptual structure for every set of them.

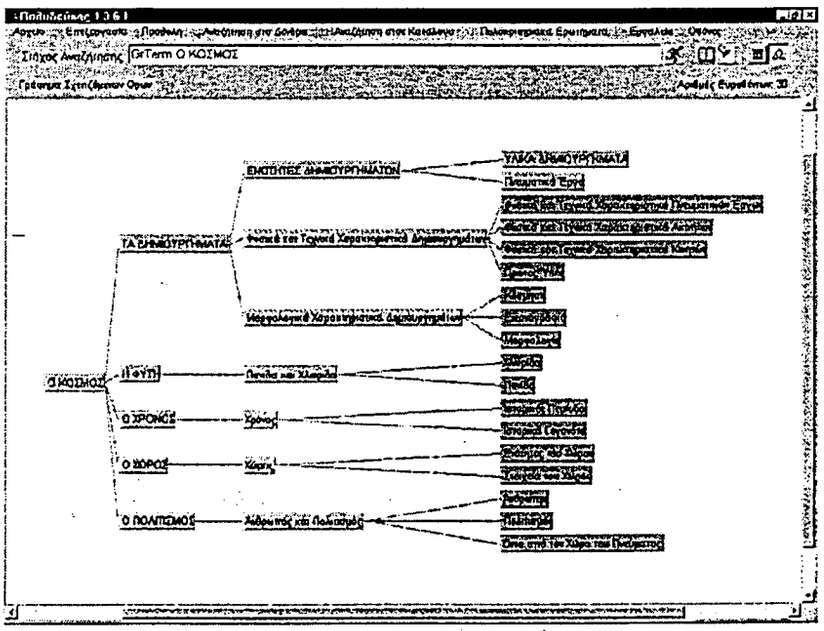


Figure 3: Screen dump from Polydeykis, "The Kosmos"

Let us illustrate the above by the example the concept "temple" (in Greek "Ναός"). This concept has two designated uses in Polydeukis: a descriptive and a characteristic historical one. The descriptive concept "temple" designates every single building of religious function, regardless of the religion or the religious tradition it belongs to. As such, it is a specialization of a morphological descriptive term "building", and the functional term "object from the religious life". "Temple" is further specialized into characteristic historic terms by the particular religious tradition it belongs to, i.e., "sanctuary, temenos, temple, dominicium, mosque" (temple appears again in its narrower sense). Associated are characteristic concepts of "elements" of temples, functional parts like: "altar, nave, opisthodomos, diakonikon, prothesis, choir, narthex, royal doors, mihrab, minbar" and structural parts like: "capital, architrave, chancel barrier, frieze, pediment, mukarnas, etc.". In a similar way, each characteristic concept is embedded in a conceptual structure capturing its social, functional, morphological and stylistic significance with the corresponding terminology.

The conceptual structure underlying all terms contained in Polydeukis follows as well the above-mentioned anthropocentric model. The distinctions of concepts in Polydeukis are purely methodological: they are meant to convey not the levels of the very nature of things, but the levels of man's cognition according to the specific

standpoint by which he searches and understands the nature of things. For that purpose, Polydeukis uses the following major facets (see figure 3):

- Kosmos , the world as subject
- Living Nature, as historical subject
- Culture and Civilization
- Space
- Time
- Creations, the man-made world:
  - Material creations
  - Conceptual works
  - Associated concepts: Stylistic, physical and technical characteristics

Hereby "Kosmos" includes the other facets as subjects of representations. The "Living Nature" is organized by the relation to the man, habitat and feeding habits, rather than genealogy. The man-made objects (as formal aspect "είδος"), their associated concepts (constitutive aspect) and relevant social/functional terms from the Culture and Civilization facet are combined systematically between each other in two levels of hierarchy of descriptive terms into a kind of regular grid. The combinations are announced groupwise by guide terms like "functional parts of immobile objects". Under these, the characteristic terms are introduced, and the hierarchy continues on demand with the multiple criteria. E.g. "immobile objects" are split into "simple constructions", "buildings", "built complexes", "housing areas", "open spaces", "other immobile objects", whereas activities are split into "religious life", "burial rites and functions", "private life" etc. (see figure 4). The combination of "buildings" and "religious life" leads e.g. to "temple", as seen above.

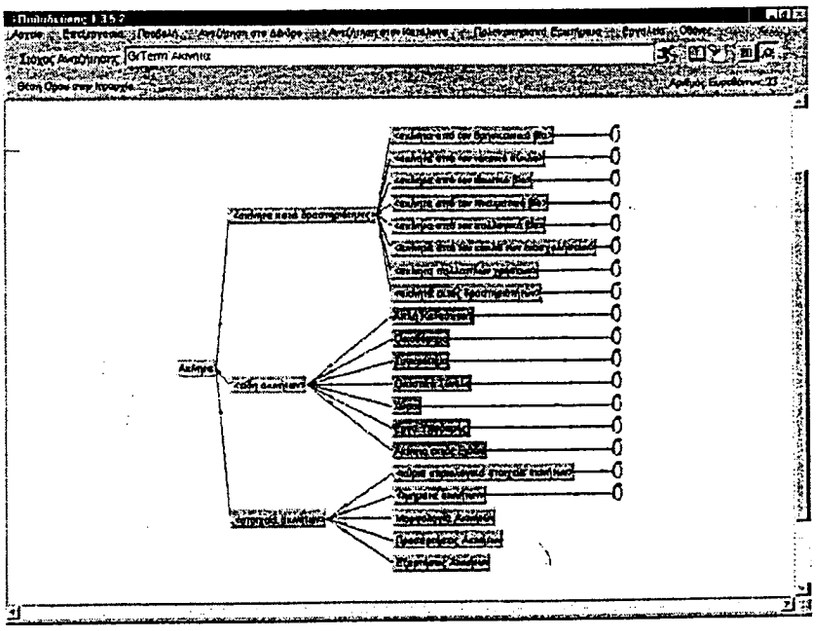


Figure 4: Screen dump from Polydeykis, immobile objects hierarchy

This kind of hierarchical layout allows for the formation of a common structural environment for the descriptive and the historical/characteristic level. According to such an organizing principle of the Polydeukis material, the higher we are in the hierarchy of terms, the more we function on a level of descriptive categorization. On the other hand, the more we descend from top terms, the more we are apt to encounter characteristic historic terms. The main characteristic of this superintending view of facets is that the underlying synthetic structuring of terms follows a flexible policy of conceptual integration which is capable to registering the whole spectrum of variegations of concepts whether descriptive or historic/characteristic. Furthermore, it allows for the formation of broader or differentiated descriptive categorizations for the same characteristic concepts.

The first experience with users is quite encouraging. The systematic structure of the upper levels and the explicit relation to descriptive concepts provide a remarkable orientation to the user, and we got enough positive responses. Characteristic terms seem to "find a natural place" in the hierarchy. On the other side, the creation of combinations leads to a kind of "combinatorial explosion", and is not easy to be maintained. As well, the introduction of guide terms at all levels makes the presentation "heavy". Speaking with Ranganathan, we have

successfully ignored the limitations at the notational plane, overcome the verbal plane by introducing compound terms, and satisfied the idea plane. The next step is now to introduce a suitable notation, and to improve compound term creation.

7 Conclusions

The experience with Polydeykis, the analysis of AAT guide terms and the linguistic theory of qualia structures make us confident, that an analysis of terms of a thesaurus about objects into a limited number of aspects and the respective logical and notational organization of thesauri is feasible and can improve considerably user orientation, precision of classification and reasoning during retrieval. We have introduced the notion of "descriptive" versus "characteristic" concept, which seems to be a viable distinction to account for certain phenomena in classification. An initial distinction into "modern" and "historical" was soon abandoned, as we could find all kinds of transitions between historical and modern terms with respect to their significance. We would be interested to learn if similar notions have been developed in philosophy or cognitive science.

We have hereby encountered limitations of the current notational practice. We believe, that these can be overcome by mechanisms of virtual precoordination based on Description Logic we have made first experiments with [Ntoas99]. Eventually also new modes of graphical presentations need to be developed. In the sequence, there are several more open questions about thesaurus structuring, like:

1) Should there be a relation between certain qualia and the level of abstraction, or between descriptive and characteristic terms in a practical thesaurus? E.g. do functional terms tend to be more abstract than characteristic or formal ones as psychological studies suggest?
2) Are the qualia of common sense language the same with those of a specialist vocabulary? Are the latter similar, a part of the latter or can they be completely alien, e.g. in some technical fields?
3) A formalization of multi-typed hyponymy for thesauri, i.e. broader / narrower term relations specific to qualia, should be developed.
4) To which detail sense distinction in thesauri should be made in order to account for historic concept shifts? Where is the optimal precision/ complexity

Finally, this approach may be useful to merge or make the transition easier between multiple thesauri in networked environments, as it can give a better account for reasons and aspects of broader/ narrower term distinctions.

8 Literature

BECHHOFER, S., GOBLE, C.A, 1999. *Classification Based Navigation and Retrieval for Picture Archives.* IFIP WG2.6 Conference on Data Semantics, DS8 , Rotorua, New Zealand.

BORGIDA, A., 1995. *Description Logics in Data Management.* IEEE Transactions on Knowledge and Data Engineering, 7(5):671-682.

CONSTANTOPOULOS P. AND DOERR M., 1995. *Component Classification in the Software Information Base.* O. Nierstrasz and D. Tsichritzis, eds. Object-Oriented Software Composition. Prentice-Hall.

CHEN, H., LYNCH K., BASU K. AND NG, D., 1993. *Generating, Integrating, and Activating Thesauri for Concept-based Document retrieval.* IEEE Expert. 8(2):25-34.

[DMOZ] DMOZ Open Directory Project http://dmoz.org/

[DL] More information can be found for instance on http://www.ida.liu.se/labs/iislab/people/patla/DL/index.html

DOERR, M AND FUNDULAKI, I, 1998. *A proposal on extended interthesaurus links semantics.* Heraklion - Crete, Greece: FORTH, Institute of Computer Science -Technical Report FORTH-ICS/TR-215.

DOERR, MARTIN AND CROFTS, NICHOLAS, 1999. *Electronic Esperanto: The Role of the Object Oriented CIDOC Reference Model.* Proc. of the ICHIM'99, Washington, DC.

[Encarta] Microsoft Corporation: "Mircrosoft Encarta Reference Suite 2000 , CD SET X04-99037, 1993-1999

HAZEWINKEL, MICHIEL, 1996. *Enriched Thesauri and their Uses in Information Retrieval and Storage.* FIRST DELOS WORKSHOP, An Overview on Projects and Research Activities in Digital Library Related Fields". INRIA - Sophia Antipolis. ERCIM Workshop Proceedings - No. 97-W001.

[MDA97] Museum Documentation Association, English Heritage & Royal Commission on the Historical Monuments of England, "mda Archaeological Objects Thesaurus", 1997, http://www.open.gov.uk/mdocassn/archobj/archcon.htm

MICHALSKI, R.S., *Beyond Prototypes and Frames: The Two-tiered Concept Representation.* R. S. Michalski in Categories and Concepts, Theoretical Views and Inductive Data Analysis editors I. Mechelen, J. Hampton, R. Michalski and P. Theuns.

MILLER, A. GEORGE, BECKWITH, RICHARD, FELLBAUM, CHRISTIANE, GROSS, DEREK AND MILLER, KATHERINE, 1993. *Introduction to WordNet: An On-Line Lexical Database.*

MURRAY-RUST, PETER AND WEST, LESLEY, 1998. *Knowledge, Language and Semantics: XML and VHG (TM).* ASLIB Managing Information vol.5, no.4, pages 34-36.

NTOAS, DIMITRIS, 1999. *Economy and consistency in Thesauri.* Heraklion - Crete, Greece: FORTH, Institute of Computer Science -Technical Report FORTH-ICS-TR-262.

PINKER, STEVEN, 1994. *The Language Instinct.* New York: W.Morrow and Co.

POLLITT, STEVEN, 1997. *Interactive Information Retrieval based on Faceted Classification using Views- Knowledge Organization for Information Retrieval.* Proceedings of the 6th International Study Conference on Classification, University College, London 16-19 June 1997. http://www.hud.ac.uk/schools/cedar/dorking.htm

POLLITT, STEVEN, 1997. *The key role of classification and indexing in view-based searching.* IFLA '97 Copenhagen Aug 31 - Sept 3 1997. 63rd IFLA General Conference Booklet 4. Section on Classification and Indexing Session 95 Paper 009-CLASS-1-E .

PRIETO-DIAZ, R., AND FREEMAN, P., 1987. *Classifying Software for Reusability*, IEEE Software. pp.6-16.

[PrivCom] Private Communication with Tanya Szraiber, British Museum.

PUSTEJOVSKY, JAMES, 1995. *The Generative Lexicon*. MIT Press. ISBN 0-262-16158-3.

RANGANATHAN, S.R., 1965. *A descriptive account of Colon Classification*. Bangalore: Sarada Ranganathan Endowment for Library Science.

RECTOR, A., BECHHOFER S., GOBLE, C., HORROCKS I., NOWLAN, W., AND SOLOMON, W., 1997. *The GRAIL Concept Modelling Language for Medical Terminology*. THE GALEN Project, Medical Informatics Group, Department of Computer Science, University of Manchester, England.

[SHIC93] The SHIC Working Party, " Social, Historical and Industrial Classification", The Museum Documentation Association, Cambridge 1993, ISBN 0 905963 91 1

SVENONIUS, ELAINE, 1989. *Design of Controlled Vocabularies*. Encyclopedia of Library and Information Science, New York, Marcel Dekker, p82-109.

[TEI] Text Encoding Initiative (http://www.uic.edu/orgs/tei), .Chapter 13, "Terminological Databases" http://www.uic.edu/orgs/tei/p3/doc/p3te.txt.

VICKERY, B.C., 1960. *Thesaurus - A new word in documentation.* Journal of Documentation, 16:4, pp. 181-89.

WELTY, CHRIS, AND JENKINS, JESSICA, 1999. Welty, *Formal Ontology for Subject*. J. Knowledge and Data Engineering. 31(2)155-182. Elsevier.

WELTY, CHRIS, AND IDE, NANCY, 1999. *Using the Right Tools: Enhancing Retrieval from Marked-up Documents*. Journal of Computers and the Humanities. 33(1-2). Kluwer

SOERGEL, D., 1995. *The art and architecture thesaurus (AAT): A critical appraisal.* Visual Resources, X, 369-400.

Metadata? Thesauri? Taxonomies? Topic Maps!

Making sense of it all

Lars Marius Garshol <larsga@ontopia.net>

Abstract

To be faced with a document collection and not to be able to find the information you know exists somewhere within it is a problem as old as the existence of document collections. Information Architecture is the discipline dealing with the modern version of this problem: how to organize web sites so that users actually can find what they are looking for.

Information architects have so far applied known and well-tried tools from library science to solve this problem, and now topic maps are sailing up as another potential tool for information architects. This raises the question of how topic maps compare with the traditional solutions, and that is the question this paper attempts to address.

The paper argues that topic maps go beyond the traditional solutions in the sense that it provides a framework within which they can be represented as they are, but also extended in ways which significantly improve information retrieval.

Table of Contents

1. Introduction 2
2. Objects and their metadata 2
    2.1. What is metadata? 2
    2.2. Metadata as a finding aid 2
    2.3. Subjects and precision 3
3. Subject-based classification 4
    3.1. Controlled vocabularies 5
    3.2. Taxonomies 5
    3.3. Thesauri 7
    3.4. Faceted classification 8
    3.5. Ontologies 8
    3.6. Other subject-based techniques 9
4. Topic maps 9
    4.1. The names of subjects 10
    4.2. Types 10
    4.3. Occurrences 11
    4.4. Associations 11
5. Comparison 12
    5.1. Traditional classifications in topic maps 12
    5.2. Merging metadata and classification 13
    5.3. Benefits and costs 14
    5.4. Searching 15
    5.5. Schemas 15
    5.6. Identity and merging 16
6. Conclusion 16
Bibliography 17
Glossary 17

Rendered by www.RenderX.com

1. Introduction

The task of an information architect is to create web sites where users can actually find the information they are looking for. As the ocean of information rises and leaves what we seek ever more deeply buried in what we don't seek, this discipline becomes ever more relevant. Information architecture involves many different aspects of web site creation and organization, but its principal tools are information organization techniques developed in other disciplines. Most of these techniques come from library science, such as thesauri, taxonomies, and faceted classification.

Topic maps are a relative newcomer to this area and bring with them the promise of better-organized web sites, compared to what is possible with existing techniques. However, it is not generally understood how topic maps relate to the traditional techniques, and what advantages and disadvantages they have, compared to these techniques. The aim of this paper is to help build a better understanding of these issues.

2. Objects and their metadata

Metadata is the foundation of all information retrieval, and so we start out by examining what metadata really is. A later section will consider the relation between metadata and library science techniques.

2.1. What is metadata?

It is generally assumed when organizing information that it consists of discrete pieces, though the terms used for these vary. The pieces are sometimes referred to as "documents", at other times as "objects". We will use the term *object* here for the entities being organized, as it does not seem appropriate to assume that they will all be documents in the traditional sense of the word.

Metadata is generally defined as "data about data," which is of course a very broad definition. In computer science this is generally taken to mean information about a set of data in a particular representation, which typically means schema information, administrative information, and so on. However, in content management and information architecture, metadata generally means "information about objects" ("objects" here used as defined above), that is, information about a document, an image, a reusable content module, and so on. Since it is the management of content we are primarily concerned with here, this is the definition we will use throughout this paper.

The best-known vocabulary for metadata is Dublin Core [Dublin Core], which is a set of 13 properties that may be applied to information resources to describe them. These properties contain information such as "title", "creator", "subject", "description", "publisher", "date", "language", etc. The Dublin Core specification defines the meaning of each property, but is silent on how to represent both the properties and their values, and is thus independent of any particular technology. Dublin Core was intended to aid "resource discovery", that is, it was meant to support information retrieval. However, it's now commonly agreed that the metadata is as useful for the management of content as it is for the discovery of it after publication, and so metadata in practice tends to be used for both purposes.

In general, metadata is best understood as "any statement about an information resource", regardless of what it is being used for, which metadata vocabulary is being used, and how the metadata is represented. In this paper we will focus on the use of metadata as a finding aid, and ignore the other uses to which it may legitimately be put.

2.2. Metadata as a finding aid

Obviously, searching for specific information in a large document corpus in the absence of *any* form of metadata (that is, information about the objects) is pretty much a hopeless task. The question is, what kind of information about the objects would help the user the most?

One common case is where the user has seen the object that is being sought once already, and so the user may remember specific details about it, such as words from the title, who wrote it, or when it was written. These clues can then be used to find the document by searching using the clues and trying out different searches until the right document shows up. Dublin Core metadata supports this form of retrieval quite well, since this is precisely the sort of information it contains.

Rendered by www.RenderX.com

This is a special case, however, and in the more general case the user wants information about a specific subject[1], and sits down in front of whatever user interface is offered to find the answer to the question "what objects are about subject X?" The question then is, what help can the user interface give this user?

If we assume that the interface is based on Dublin Core metadata, it turns out that the answer is: not that much. Below is shown Dublin Core metadata for a paper presented at XML Europe 2003.

Title	Curing the Web's Identity Crisis
Creator	Steve Pepper & Sylvia Schwab
Subject	RDF, topic maps, subject indicator
Publisher	IDEAlliance
Date	May 2003
Language	English
Format	XML

Table 1.

Clearly this information does not help the reader much in trying to establish what the paper is actually about, though from the subject we can see that it has something to do with topic maps and RDF. This highlights the problem: that standard metadata mostly provides administrative information, and that it says very little about the subject of an object. Of the Dublin Core metadata properties only a few address this question, and most of them do so indirectly:

title — The title of the document usually offers good clues as to what the document is about, but it does not necessarily mention all names of all subjects the user is interested in, and it may also presuppose knowledge the user does not actually possess.

description — This field is likely to describe what the document is about, but again may not facilitate search and discovery very effectively, for the same reasons that the "title" field may fail to do so.

subject — This field, which usually contains a set of keywords, is meant to convey precisely what the document is about. However, much depends on how extensive the set of keywords is, whether all related subjects are mentioned, and whether too many subjects are listed, leading the user to get too many hits.

2.3. Subjects and precision

In addition to metadata not necessarily saying very much about the aboutness of an object a related problem is that making metadata describe the subject precisely may also be difficult. Let's imagine a user sitting down in front of an interface to all the papers presented at the IDEAlliance conferences, using a search interface based on Dublin Core, and looking for information on topic maps. The user is new to topic maps, and not yet interested in any specific aspect of topic maps, just the subject area in general.

If the user now does a search for "topic maps" in the keywords (that is, the "subject" field), all papers which mention "topic maps" as a keyword will be in the search result. One problem is that this will tend to be both introductory material as well as more specialized material, and the results will be a simple list of documents, probably showing title, author, and date.

The title, author, date, and description fields are very useful here, as they help the user choose between the results, but what of the quality of the results themselves? I actually tried this out on the document corpus described in [The

[1] The relation between the word 'subject' as used here and as defined in topic maps is important, and we will return to it further down. Meanwhile, take 'subject' to mean 'any concept in which the user may potentially be interested'.

Rendered by www.RenderX.com

XML Papers] [2], and got the result shown below, in "most recent first" order. (A metadata structure as simple as Dublin Core does not allow a "most relevant" ranking.)

Title	Creator	Date
Tax Map: An integrated navigation tool for the IRS Call Center Research System	Michel Biezunski	December 2003
Curing the Web's Identity Crisis: Subject Indicators for RDF	Steve Pepper & Sylvia Schwab	December 2003
OKS 2.0: New Utensils for Topic Map Chefs	Pamela Gennusa	December 2003
Semantic Web Servers - Engineering the Semantic Web	Graham Moore	December 2003
Information Architecture with XML	Peter Brown	December 2003
BookBuilder: Content Repurposing with Topic Maps	Nikita Ogievetsky & Robert Sperberg	December 2003
The TAO of Topic Maps	Steve Pepper	December 2003

Table 2.

A glance at this list shows that most of these documents are not primarily about topic maps, but about subjects *related to* topic maps. However, if the authors had not listed "topic maps" as a keyword those searching for "topic maps" would have been unable to find their papers at all.

Another problem with this corpus is that authors have been required to define their own keywords, which means that the choice of keywords can be quite eclectic. A random pick of the more unusual keywords from the corpus mentioned above finds "'xml' reserved", "analyze eval approach", "pool", "fnctional composition" [sic!], "semantic", "Contrasting approaches: DTD, W3C schema, RELAX NG", "Exposition: Recent history of MPEG-7", etc. Clearly these are not good keywords, for several reasons.

Different forms of the same keyword, or closely related keywords, is also a problem. In the corpus can be found "topic navigation maps" (old name for topic maps), "topic maps", "XML topic maps" (a format for topic maps, often used as a synonym), "XTM" (the acronym for the former), and so on. The problem here is that four keywords refer to two subjects, and that the two subjects are very closely related. None of this is actually captured, and the user simply has to search to find out, or know beforehand.

To conclude: the most useful metadata about a document is the keywords, since that is the only thing that explicitly describes what the document is about. The other metadata are useful in managing the documents and in helping the user decide which of their search hits they want to look more closely at.

The other conclusion to be drawn from this analysis is that having keywords be just a simple text field with no restrictions is not going to work very well.

3. Subject-based classification

Subject-based classification is any form of content classification that groups objects by the subjects they are about. This can take many forms, and is generally combined with other techniques in order to create a complete solution. In the example above, the use of keywords to classify papers is a subject-based classification approach.

The relation between subject-based classification and metadata is that metadata properties or fields that directly describe what the objects are about by listing discrete subjects use a subject-based classification. This basic feature is common to all subject-based classifications, and as we will see the differences between the various techniques lie in what they say about the subjects, rather than in what they say about the objects.

[2]The actual query, for those who would like to know, was:
discussed-in(topic-maps : keyword, $PAPER : paper), presented-at($PAPER : paper, $CONFERENCE : conference), when-event($CONFERENCE : event, $DATE : time) order by $DATE desc?

Rendered by www.RenderX.com

This needs to be emphasized: there is a difference between describing the objects being classified and describing the subjects used to classify them. What we will discuss in this section are the different approaches to describing subjects. Metadata describes objects, and one of the ways in which it does that is by connecting objects to the subjects they are about. We will return to this idea below.

3.1. Controlled vocabularies

Controlled vocabulary is a rather broad term, but here we mean by it a closed list of named subjects, which can be used for classification. In library science this is sometimes known as an *indexing language*. The constituents of a controlled vocabulary are usually known as terms, where a *term* is a particular name for a particular concept. (This is pretty much the same as the common-sense notion of a keyword.)

It is common to distinguish between *term* and *concept* by saying that the former is the name of a concept, and that the same concept may have multiple names, and also that the same term may name multiple subjects. A controlled vocabulary consists of terms, and not directly of concepts, and in general each term will be disambiguated to refer to a single subject (that is, there will be no duplicate terms). Note that "subject" as we have used the term so far is effectively equivalent to "concept".

The term vocabulary also means slightly different things in the term "controlled vocabulary" from what it does in "metadata vocabulary". The former is, as we noted, a set of indexing terms, or subjects used for classification, while the second is a set of properties of objects.

The purpose of controlling vocabulary is to avoid authors defining meaningless terms, terms which are too broad, or terms which are too narrow, and to prevent different authors from misspelling and choosing slightly different forms of the same term. Thus we can prevent authors from using "topic navigation maps" and "topic map" by forcing them to choose "topic maps". We can also make it impossible to use "fnctional composition" as an indexing term instead of the correct "functional composition".

The approach taken for IDEAlliance conference proceedings is what is known as an *uncontrolled vocabulary*, and this was abandoned recently in favour of a controlled vocabulary for precisely the reasons cited above. The simplest form of controlled vocabulary is simply a list of terms and nothing more. This is the approach currently used for the IDEAlliance conference proceedings, but more advanced schemes exist, as we will see below.

3.2. Taxonomies

The term taxonomy has been widely used and abused to the point that when something is referred to as a taxonomy it can be just about anything, though usually it will mean some sort of abstract structure. Taxonomies have their beginning with Carl von Linné[3], who developed a hierarchical classification system for life forms in the 18th century which is the basis for the modern zoological and botanical classification and naming system for species. In this paper we will use *taxonomy* to mean a subject-based classification that arranges the terms in the controlled vocabulary into a hierarchy without doing anything further, though in real life you will find the term "taxonomy" applied to more complex structures as well.

The benefit of this approach is that it allows related terms to be grouped together and categorized in ways that make it easier to find the correct term to use whether for searching or to describe an object. For example, this could help users and authors by making it clear that there are two closely related terms: "topic maps" and "XTM", and helping them choose the right one. (Or, at least, for the users, telling them that they should perhaps try both.)

[3]Originally Carl Linnæus. The traditional English misspelling of the name is Carl Linnaeus.

Rendered by www.RenderX.com

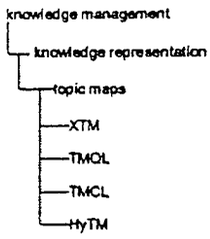


Figure 1.

The figure above shows the placement of topic maps within a hypothetical taxonomical structure. As can be seen, this structure could easily help someone looking for information on topic maps or classifying a document to do with topic maps to pick the right terms to use.

Note that the taxonomy helps users by describing the *subjects*; from the point of view of metadata there is really no difference between a simple controlled vocabulary and a taxonomy. The metadata only relates objects to subjects, whereas here we have arranged the subjects in a hierarchy. So a taxonomy describes the subjects being used for classification, but is not itself metadata; it can be used in metadata, however. The diagram below illustrates this.

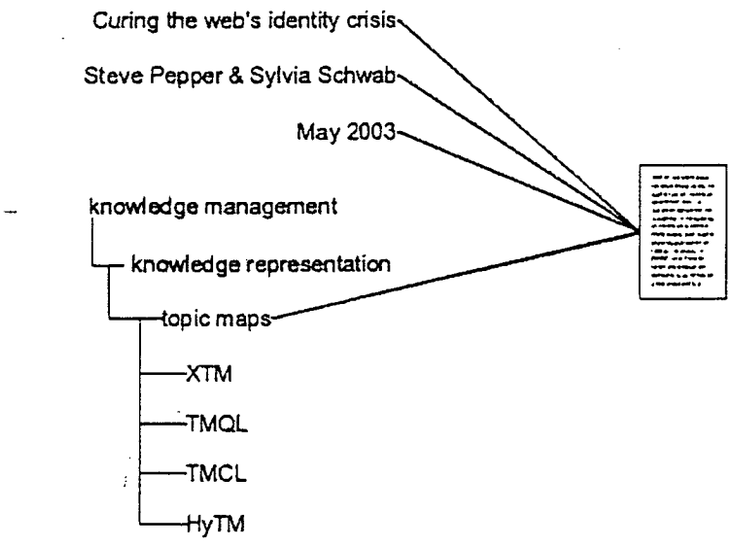


Figure 2.

In this diagram, the blue lines are the metadata, while the black lines that make up the taxonomy is part of the subject-based classification scheme. The distinction derives from the blue lines being statements about the paper, but the black line between "topic maps" and "knowledge representation" is *not* a statement about the paper; it's a statement about "topic maps". One consequence of this is that if we have another paper about "topic maps" we do not need to repeat that "topic maps" belong under "knowledge representation".

As we said, the taxonomy provides more information about the concepts, and it does so to help the users. However, while the taxonomy does help the user, a number of important pieces of information about the concepts are not being captured here, such as:

- The fact that "XML Topic Maps" is synonymous with "XTM".
- The difference between "XTM" and "topic maps". (Many users use these interchangeably, but they do not mean the same thing.)
- The fact that "topic navigation maps" is synonymous with "topic maps", but should no longer be used.
- The relationship between topic maps and subject-based classification and topic maps and the semantic web.

Rendered by www.RenderX.com

- The relationship between XTM and XML and HyTM and SGML.
- The similiarity between HyTM and XTM, and their difference from TMQL and TMCL, as well as the similarity between TMQL and XQuery.

All of these have consequences for the end user, since it means that they must search using precisely the right term, look in precisely the right places to find the terms, and so on. A taxonomy as we defined it here cannot handle these problems, though it should be noted that many systems referred to as taxonomies to some extent can, as they extend the basic model defined here.

3.3. Thesauri

Like the term "taxonomy" the term "thesaurus" has been used to describe all kinds of subject classification structures, though for thesauri there are two ISO standards describing their structure. [ISO2788] describes monolingual thesauri, while [ISO5964] is for multilingual thesauri. We will here discuss thesauri as they are defined in the ISO standards, while noting that in practice many users extend the structure somewhat, and in some cases the term is applied to structures differing substantially from what is described here.

Thesauri basically take taxonomies as described above and extend them to make them better able to describe the world by not only allowing subjects to be arranged in a hierarchy, but also allowing other statements to be made about the subjects. [ISO2788] provides the following properties for describing subjects.

B Short for "broader term", refers to the term above this one in the hierarchy; that term must have a wider or less specific meaning. In practice some systems allow multiple BTs for one term, while others do not. (There exists an inverse property known as NT, for "narrower term", which is implied by the BT.)

One could say that taxonomies as described above are thesauri that only use the BT/NT properties to build a hierarhcy, and don't make use of any of the properties described below, so it could be said that every thesaurus contains a taxonomy.

S This is a string attached to the term explaining its meaning within the thesaurus. This can be useful in cases where the precise meaning of the term is not obvious from context. "SN" stands for "scope note".

Since users often use "XTM" when they mean "topic maps", it would be useful to add a scope note to XTM saying something like "The standard XML interchange format for topic maps. When discussing topic maps in general, and not just the format specifically, use the term 'topic maps'."

E Refers to another term that is to be preferred instead of this term; implies that the terms are synonymous. (There exists an inverse property known as UF.)

For example, on "topic navigation maps" we could put a "USE" property referring to "topic maps". This would mean that we recognize the term "topic navigation maps", but that "topic maps" means the same thing, and we encourage the use of "topic maps" instead.

If we do this we would also have a "UF" property on "topic maps" referring to "topic navigation maps", since this is implied by the "USE" relationship.

T Is short for "top term", and refers to the topmost ancestor of this term. The term at the other end of this property is the one that would be found by following the "BT" property until you reach a term that has no "BT". This property is strictly speaking redundant, in the sense that it doesn't add any information, though it may be convenient.

R Short for "related term", refers to a term that is related to this term, without being a synonym of it or a broader/narrower term.

For "topic maps" we could use this to indicate that "subject-based classification" and "ontologies" are terms related to "topic maps".

Rendered by www.RenderX.com

In short, thesauri provide a much richer vocabulary for describing the terms than taxonomies do, and so are much more powerful tools. As can be seen, using a thesaurus instead of a taxonomy would solve several practical problems in classifying objects and also in searching for them.

3.4. Faceted classification

The term *faceted classification* has (you saw this coming, right?) been used to mean many different things. It was originally proposed by S.R. Ranganathan in the 1930s[Ranganathan], and works by identifying a number of *facets* into which the terms are divided. The facets can be thought of as different axes along which documents can be classified, and each facet contains a number of terms. How the terms within each facet are described varies, though in general a thesaurus-like structure is used, and usually a term is only allowed to belong to a single facet [Svenonius].

In faceted classification the idea is to classify documents by picking one term from each facet to describe the document along all the different axes. This would then describe the document from many different perspectives. Ranganathan's original proposal (also known as Colon Classification) consisted of five facets:

Personality	This facet was intended for the primary subject of the document, and is considered the main facet.
Matter	The material or substance the document deals with.
Energy	The processes or activities the document describes.
Space	The locations described by the document.
Time	The time period described by the document.

The classification of a book on Norwegian rural architecture in the 17th century might run something like this:

- Personality — architecture
- Matter — wood
- Energy — design
- Space — Norway
- Time — 17th century

Faceted classification may seem very different from a thesaurus, but in fact faceted classification could be seen as simply a very disciplined way to construct a thesaurus as well as to use it for classification purposes. This is not to say that faceted classification is not useful, only that it is less different from thesauri than it may seem at first glance.

There exists an XML interchange syntax for faceted classification, known as eXchangeable Faceted Metadata Language (TMCL), which was inspired by XTM, and has some features in common with it. XFML does not require the use of any specific set of facets, nor any specific set of terms within each facet, but does use a thesaurus-like structure for the terms within each facet. For more information, see [van Dijck].

It should be noted that there exists a generalized view of faceted classification wherein each facet is generalized to the point where it becomes a general property, and where the notion of a document is generalized to mean any kind of object. In this view of faceted classification there is little difference between faceted classification and ontologies as they are described below.

3.5. Ontologies

The term ontology has, need we say it, been applied in many different ways, but the core meaning within computer science is a model for describing the world that consists of a set of types, properties, and relationship types. Exactly

Rendered by www.RenderX.com

what is provided around this varies, but this is the essentials of an ontology. There is also generally an expectation that there be a close resemblance between the real world and the features of the model in an ontology.

In this section we have essentially been discussing languages for describing the subjects used in subject-based classification, steadily progressing towards more powerful means of description. Ontologies represent the culmination of this progression, in the sense that all of the above are fixed-vocabulary languages for subject description, while ontologies have open vocabularies.

In a taxonomy the means for subject description consist of essentially one relationship: the broader/narrower relationship used to build the hierarchy. The set of terms being described is of course open, but the language used to describe them is closed, since it consists only of a single relationship.

Thesauri extend this with the RT and UF/USE relationships, and the SN property, which allow them to better describe the terms. (TT, being redundant, is ignored here.) Again the language is closed, since this is the entire vocabulary at disposal for describing the terms. In fact, thesauri could in theory be considered an ontology where there is only one type, called "term", one property, called "scope note", and three relationships (BT/NT, USE/UF, and RT). In practice thesauri are not considered ontologies because their descriptive power is far too weak, precisely because of this limited vocabulary.

Faceted classification does not really extend this language, but provides a consistent and useful discipline for applying it. That is, faceted classification does not introduce any new properties or relationships, though it could be said to have a new type: facet. It could be described as simply requiring the creator of the indexing language to create a set of facets and then fill each with a thesaurus that does not overlap with the others.

With ontologies the creator of the subject description language is allowed to define the language at will. Ontologies in computer science came out of artificial intelligence, and have generally been closely associated with logical inferencing and similar techniques, but have recently begun to be applied to information retrieval. A number of technologies and tools exist in this area, but we will focus on topic maps in this paper, since they were created to be an ontology framework for information retrieval.

3.6. Other subject-based techniques

A number of other terms and techniques are commonly applied today within information architecture, without fitting neatly into the classification scheme used in this section. One such term is *categories*, which are often used to group objects on web sites. Categories are just terms in a subject-based classification, that is, a controlled vocabulary. The categories can be a plain list, or they can be arranged in a taxonomy. That's really all there is to it. (This works the other way as well; every taxonomy or thesaurus consists of categories into which objects are grouped.)

Another technique that is much used is *synonym rings*, which connect together a set of terms as being equivalent for search purposes. (That is, if you search for "topic navigation maps" you should also find "topic maps", for example.) Essentially synonym rings express a synonym relationship between a set of terms, and so is similar to the UF/USE relationship of thesauri, except that there is no indication of one term being preferred above the others. Synonym rings are a rather special-purpose construct, but their function, to express that a set of terms are synonymous (that is, refer to the same concept), is very much worthwhile. The term "synonym ring" alludes to the fact that within a synonym ring every term is synonymous to every other term in the same ring; mathematicians know this as an *equivalence class*.

An *authority file* is similar to a synonym ring, the only difference being that it consists of UF/USE relationships instead of synonym relationships. So in an authority file one term in each synonym ring is indicated as being the preferred term for that subject.

What this means is that a thesaurus includes not just a taxonomy, but also an authority file. It does not include a synonym ring, however, though it could be used to support searching in the same way that a synonym ring is used.

4. Topic maps

Topic maps originated in work on the merging of electronic indexes and so are very much a subject-based classification technique. In fact, topic maps are organized around *topics*, and each topic is used to represent some real-

Rendered by www.RenderX.com

world thing. In the terminology we used above, topics represent concepts, the same way terms in an indexing language refer to concepts. In topic maps the concepts are called *subjects*, and the standard emphatically states that a subject can be "anything whatsoever". We will return to the consequences of this later on.

In topic maps, three constructs are provided for describing the subjects represented by the topics: names, occurrences, and associations. These describe the names, properties, and relationships of subjects, respectively, and we will cover the uses of each of these in more depth in the following sections. We will not give an in-depth introduction to topic maps, as that is beyond the scope of this paper. Those wanting more information should read the classic introduction [Pepper].

4.1. The names of subjects

In topic maps, a topic can be given any number of names. Giving more than one name to a topic effectively means that all the names refer to the same subject; that is, that the names are all synonyms. This means that every topic is really a synonym ring, though it can be a ring with just one member.

It is allowed in topic maps for different topics to have the same name[4], something that taxonomies and thesauri do not allow. In practice this is something that occurs all the time, and so it is important to support it. Many classification systems avoid duplicate names by including some disambiguation in the name itself. For example, Paris the city and Paris the hero of Greek mythology may have been differentiated as "Paris (France)" and "Paris (Greek myth.)". In topic maps this is not necessary, and the types, occurrences, and associations of the topics will generally distinguish them anyway.

The power of topic maps with regards to names does not stop here, however. A name can be given a *scope*, which is a set of topics representing the context in which the name is appropriate. A name that has an empty scope is considered to have unlimited validity. This enables the creator of the topic map to define a kind of language for describing names.

For example, one could create a topic for topic maps and give it the name "topic maps" (with empty scope), but also the name "topic navigation maps", in the scope "obsolete". This would have the same effect as the USE/UF construct in thesauri, or as an authority file, but is more powerful, since one can say *why* the term "topic navigation maps" shouldn't be used. (The reason being, of course, that it is obsolete.)

This approach to names enables new usage areas that do not exist in the same way with traditional tools. For example, different terminologies are often used within a single organization, whether because of regional differences, or because of differences in corporate cultures internally in the organization. Topic maps can support this by scoping the names with topics representing the corporate cultures or areas where the terms are used. This allows the topic map to say that "corporate culture A calls this region 'APAC', while culture B calls it 'Asia-Pacific'.

Users can then in their profile (which may be permanent, or just for a single visit to the site) state their preferred terminology, and the site can then display the correct names for each topic for them. In cases where a topic has no name in the scope "corporate culture A", the unscoped name will be chosen instead.

This can also be used to support multilingual information, since topics representing languages can also be used as scopes. Thus we could give the topic for topic maps the names "emnekart" (Norwegian) and "???????" (Japanese), and have the system display the names in the user's preferred language.

Since scopes consist of topics, it is up to the creator of the topic map to define their own language for describing the names of subjects. Topic maps provide all the power of traditional techniques in this regard, but go far beyond it, in that they remove several of the restrictions, and also allow the relationship between the terms and the subjects to be described in full, instead of expressed in terms of specific desired behaviour ("prefer this term", or "treat these terms as equivalent").

4.2. Types

In topic maps topics can be typed, which provides considerable power for describing the world from which the topics are taken. This is a capability that is missing from traditional classification techniques. Using this, one could

[4] You may have heard of the "topic naming constraint", which is a rule in topic maps that only allows duplicate names in certain limited cases. This rule has now been removed from topic maps, so the statement *is* true.

Rendered by www.RenderX.com

create types and assign them to topics, and thus say that "topic maps" is-a "technology", "XTM" is-a "interchange format", "Norwegian" is-a "language", "HyTM" is-a "interchange format", "TMCL" is-a "constraint language", and "TMQL" is-a "query language", and so on. (This covers a number of the statements we couldn't express in Section 3.2)

This may sound like a very simple capability, and so it is, but it is also very powerful, and provides the foundation for several important capabilities that we will return to. The most immediate benefit, however, is that in traditional techniques the set of terms is in a sense flat, since there is no way to distinguish different *kinds* of terms. Languages are bundled in with people, places, technologies, and organizations, with no way to tell them apart[5].

Once explicit types have been provided it is possible to let the user perform searches such as "find 'paris', but show only 'places'", or to show lists of all cities, separate from other kinds of subjects. Without types there is no way to do this, since the necessary information will be missing.

Again, since the types are themselves topics the creator of the topic map can choose which types to use, which is another reason why we say that topic maps have an open vocabulary. (It may seem confusing that types are not among the constructs listed at the beginning of this section. The reason is that type assignments are considered a relationship, that is, associations, which we will return to below.)

4.3. Occurrences

Occurrences relate topics to the information they are relevant to. They effectively perform the same function as the page numbers in a back-of-book index: they indicate where information about the subject can be found in the book. And so it is with occurrences, they connect the topics to information resources that contain information about them. This is effectively the inverse of the "subject" property in Dublin Core, which connects an object to the subjects it is about. Occurrences in topic maps do the same thing, but go from the object (or information resource) to the subject (represented by a topic).

In topic maps the representation of this is more structured, however. Occurrences have types, which allow us to distinguish between different kinds of relationships to the information resource. One can distinguish a biography from a portrait, a description from a tutorial, a video clip from a specification, and so on. Scope can also be applied to an occurrence, for example to distinguish material suitable for novices from that suitable for intermediate learners.

Occurrences use URIs to identify the information resource being connected to the topic, which means that any kind of information resource anywhere can be connected to the topic. However, the occurrence does not need to use a URI; the information can also be given as a string stored directly in the topic map. This is useful for attaching simple properties to topics, such as date of birth, phone number, description, and so on.

Again, the occurrence types are topics, and so the creator of the topic map is free to define occurrence types at will. To compare with traditional techniques, they generally do not provide an open vocabulary for properties and occurrences, and although in some cases they can be extended to support it that support tends to be weak. Having very specific properties like "phone number", "date of birth" etc on terms can be very awkward when terms cannot be typed, since most specific properties only apply to a few types and make no sense when attached to other types. (What is the date of birth of a city? Or the phone number of a technology?)

4.4. Associations

Associations represent relationships between subjects, and like occurrences they can be typed. This allows any kind of relationship to be expressed. The relationships in traditional classification schemes have very little semantic content, whereas in topic maps one generally tries to make the typing of associations as specific as possible.

Associations are the final construct we need in order to be able to fully represent the set of statements given at the end of Section 3.2, and the diagram below shows the result.

[5] In faceted classification the division into facets can to some extent provide this, though it need not do so, and even when it does the types tend not to be very granular—that is, they tend to not be very precise.

Rendered by www.RenderX.com

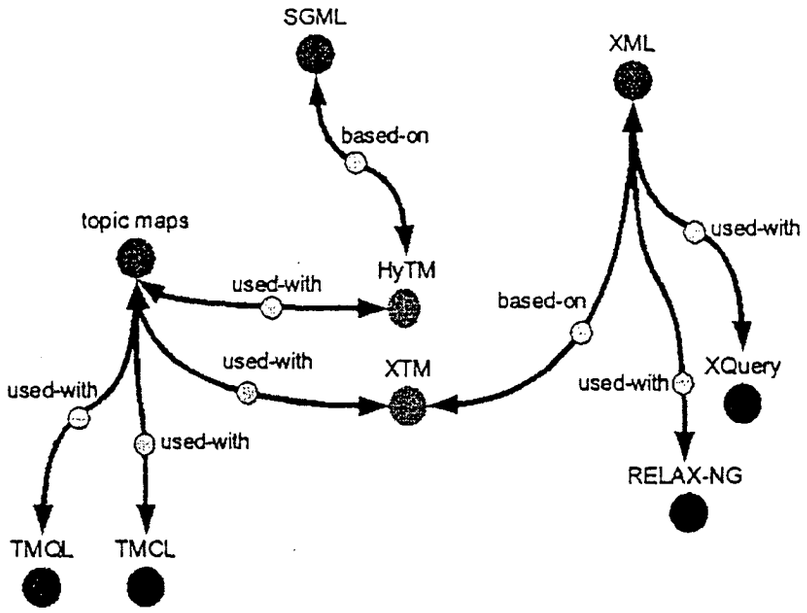


Figure 3.

In this figure, types are indicated with different colours, while we've left out the alternative names. The types of the associations can be seen from the yellow labels on the lines.

Compared to traditional classification schemes this is very different. First of all, we no longer have a hierarchy but a network of subjects. Secondly, the relationships between the subjects are clearly defined instead of being generic. From the point of view of searching, this is very powerful, since it allows us to do queries like "show me all technologies used with topic maps", or "show me every interchange format based on SGML", and so on. There are also other uses, as we will see.

5. Comparison

A summary of the relationship between topic maps and traditional classification schemes might be that topic maps are not so much an extension of the traditional schemes as on a higher level. That is, thesauri extend taxonomies, by adding more built-in relationships and properties. Topic maps do not add to a fixed vocabulary, but provide a more flexible model with an open vocabulary.

A consequence of this is that topic maps can actually represent taxonomies, thesauri, faceted classification, synonym rings, and authority files, simply by using the fixed vocabularies of these classifications as a topic map vocabulary. We'll show how this works in the following section.

5.1. Traditional classifications in topic maps

Kal Ahmed has created a standard topic map ontology that can be used to represent thesauri in topic maps [Ahmed]. There are, essentially, two ways to do this, one that models terms as subjects (one topic per term) and one that models terms as names (one name per term; one subject per concept). We will only consider the latter approach here.

To create a topic map from a thesaurus using this approach, follow these principles:

- For each term that has no USE relationship, create a topic of type "term", and make the term the name of the topic.
- For each term that has a USE relationship, make the term a a name (in scope "non-preferred term") of the topic for the preferred term.
- A scope note is an occurrence of type "scope note".

Rendered by www.RenderX.com

- The RT relationship is represented as an association of type "related term".
- The BT and NT relationships are represented as an association of type "broader/narrower" (with the roles specifying which topic is broader and which is narrower).

That's it. From the resulting topic map you can easily reproduce the original thesaurus, and the resulting structure really is a thesaurus in the same way that the original data was. Doing this is also quite easy; converting a thesaurus in a simple text format to a topic map can usually be done in a couple of hours. Note that this also covers taxonomies, since they just have names and the BT/NT relationship, so nothing more is needed to support them. Any extensions that have been made to the thesaurus or taxonomy model can easily be represented in a similar way.

For faceted classification the situation is very similar. A standard topic map ontology for faceted classification based on XFML exists, as does an XSLT stylesheet for converting XFML documents to topic maps. For more information, see [Garshol]. The XFML ontology extends the thesaurus ontology, by adding a new type "facet" and a new association type "belongs-to-facet".

Creating a topic map from a faceted classification system is thus done as follows:

- For each facet, create a topic of type "facet" and give it the name of the facet.
- For each top-level term within each facet, create a topic of type "term", with the term as the name, and associate it with the facet it belongs to using the "belongs-to-facet" association type.
- For each term below the top level, create a topic of type "term", with the term as the name, and associate it with its parent using the "broader/narrower" association type.

And that's it. You have now represented your faceted classification as a topic map.

5.2. Merging metadata and classification

The central message of this paper so far has been that metadata is statements about objects (for example, documents) while thesauri and similar techniques provide statements about subjects used in classification. Subject-based classification, of course, is the use of subjects in metadata.

Topic maps effectively unify these two approaches. Topic maps are organized around topics, which represent subjects. That is, in a topic map you find topics. Every topic you find represents a subject out in the real world that it is a symbol or stand-in for in the topic map. The definition of subject is essentially "anything whatsoever". What this means is that from the point of view of a topic map, *objects are just a special kind of subject.*

What this means is that the whole machinery we have created to describe the concepts in a subject-based classification is also available to describe the objects being classified. Thus, we can create a topic type for objects ("document", perhaps), and express the metadata using names, occurrences, and associations. The topic types let us keep track of what is a document and what is a concept, but we no longer need different technologies for metadata and classification.

And once we have everything in a single representation we can start to cross the boundaries by for example describing the authors of a paper further, and connecting them with the terms from the subject-based classification. Effectively what has happened is that the straightjacket has been removed, and we can now say anything we want to. The technology is no longer the limiting factor; instead the limits are set by our fantasy and by how much information we are able to maintain within our economic constraints.

An example of how this can be done is shown in the diagram below.

Rendered by www.RenderX.com

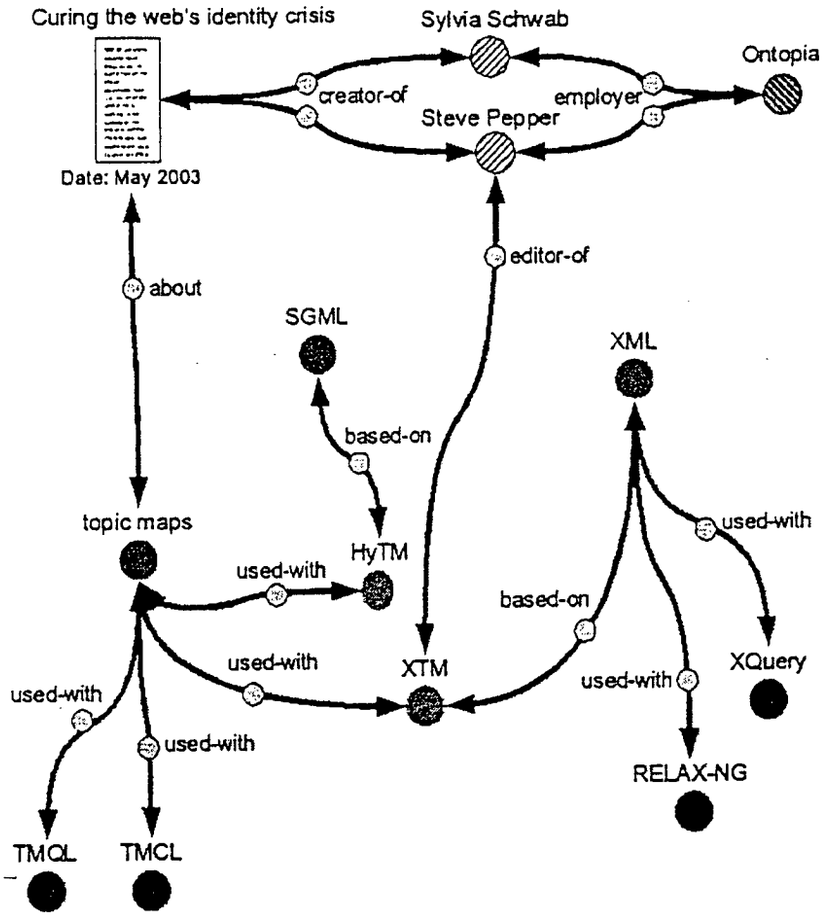


Figure 4.

Clearly, this is not doable with traditional classification systems.

5.3. Benefits and costs

The main benefit of this approach is that it leads to more richly descriptive classification and metadata systems than the traditional approaches. This again means that more precise searches are possible (see below for more on this), and that navigation systems driven by this information can be much richer and more flexible than a simple tree-based navigation.

In fact, with a richly descriptive classification system like a topic map the information in the classification system becomes a valuable resource in its own right. In many cases one will find that information that was previously provided in the classified documents, such as the structure of the organization, information about project participation and ownership etc is duplicated in the classification system. The documents are necessary, because a hierarchical classification system will not be able to express all the information one wants to capture about the organizational units and projects. With topic maps this limitation goes away, and so the documents can also go away, and their function be taken over by the navigational structure of the portal. This can effectively turn the navigational structure into a knowledge management application, while allowing it to be used for classifying content.

Quite often when using topic maps in this way one finds that having captured information in a richly structured form, this enables the creation of entirely new information products from the same source. This is similar to how moving documents from word processing formats to XML enables multi-channel publishing from the same structured source, but on a higher level.

Another benefit to using a standard like topic maps, is that a larger set of tools from various vendors become available. These tools will provide things like programming APIs, query languages, schema languages (see below),

Rendered by www.RenderX.com

graphic visualization, portal integration, content management, workflow, natural language querying, and so on and so forth.

Of course, topic maps, like any other technology, have downsides. Experience with topic maps is still in its infancy, so there may be disadvantages and risks that are currently not known. One of the more obvious are that topic maps are a relatively new technology, and so topic map expertise may be harder to find than expertise in traditional classification.

Also, while topic maps allow a richer structure and allow the users to define it themselves, this also means that the effort required to build a classification system may become greater, since there is more to build. To some extent the existence of predefined ontologies and the information design patterns provided by topic maps alleviate this, but the problem is still real, especially for newcomers to the technology.

5.4. Searching

Topic maps offer unprecedented power when it comes to searching, in the sense that they offer very good support for full-text searching, very good support for complex queries, and also provide an excellent basis for natural language querying.

Full-text searching a document corpus typically returns the documents that mentioned the term being searched for, but a topic map-based system can do far better. Instead of returning documents, the system can return the topics that best match, together with additional information, and this provides a starting point for jumping into the topic map and browsing around to find the answer to the specific question.

To take a simple example, if we were to do a full-text search for "XSLT" in the conference proceedings for the IDEAlliance conferences, there is of course a huge number of hits, but at the very top comes the topic "XSLT", which represents the XSLT standard. From there one can find the specification, papers about XSLT, which standards organization produced XSLT, tools implementing XSLT, tools using XSLT, etc

However, because of the extra structure in topic maps, we don't have to stop here. We can ask for "papers about topic map query languages", for example, or "papers about topic map query languages given by someone based in Japan", or "papers about query languages", or "papers about RDF by someone who's written papers on topic maps".

Through the use of a topic map query language these queries can get arbitrarily complex. The main problem is that the user has to know both the query language and the ontology of the topic map, and most users will be unwilling to do either. The solution to this is to build a friendly forms-based user interface where the user can set up the query by selecting terms in drop-down lists or suchlike.

There is, as mentioned above, a third possibility, which is natural language querying. Some experimentation on this has been done using topic maps, and it indicates that quite powerful search capabilities can be created by matching natural language against names in the topic map, then looking at the matched topics to see whether they are topic types, association types, or individuals, and finally putting together a query in a query language from this. It's too early to be able to conclude that this will work well for real end users, but the results so far seem very promising indeed.

5.5. Schemas

In topic maps, where users can define the classification vocabulary themselves, there is substantial risk that those creating the classification may use the vocabulary incorrectly. For example, they may use associations in ways that don't make sense, attach occurrences to topics of the wrong types, inconsistently type topics, or leave out required information.

For relational databases and XML the traditional solution to this problem is to define a *schema* that formally defines the allowed structure of the data set. This is also the solution used with topic maps, although at the moment there is no standard schema language for topic maps. A standard, called Topic Map Constraint Language (TMCL), is under development, and non-standard languages are available already.

Rendered by www.RenderX.com

Using a schema language it is possible to say things like "only persons may be employed, and they must be employed by an organization", "only persons and organizations can have phone numbers", "query languages must be used with a technology", and so on. Thus the typing capabilities of topic maps give us a starting point from which to describe the classification rules, and the schema language gives us a way to express the rules that can be enforced by software.

5.6. Identity and merging

Topic maps have a mechanism for formally declaring the identity of a topic's subject. What this means is that in topic maps there is a way to state what the subject of a topic is, in such a way that if another topic is found (for example in another topic map) that represents the same subject, one can know with certainty that it does represent *the* same subject.

The identity mechanism in topic maps is based on URIs, which can be used in two different ways. If subject of the topic is an information resource the subject can identified very easily: simply by referring to the resource using a URI. Two topics that refer to the same resource in this way must inevitably represent the same resource.

The problem is more difficult when the subject is not an information resource, but instead an abstract concept like "Norway", "Ontopia", or "Lars Marius Garshol". None of these subjects have a URI that can be used to refer to them, so the same solution cannot be used in these cases. The solution is instead to create a resource defining the subject and then referring to the defining resource.

This might lead to confusion, in that it might unclear whether the subject is the resource referred to, or the concept described by the resource. Topic maps solve this problem by distinguishing between the two kinds of references, so that it always clear what is being meant. So if two topics are found to be referring to the same resource saying "my subject is what is defined by this resource" clearly both have the same subject.

The ability to know when two topics represent the same thing provides great power, which can be exploited in many different ways:

- Topic maps from different sources which have some overlapping data can be merged together automatically (provided they have used the same identifiers).
- Topic map modules containing commonly used information can be maintained separately and reused in different contexts whenever they are needed.
- Common vocabularies can be developed as sets of defined topics and reused in different organizations. The information maintained by different users can then be merged or interchanged as desired.
- Information from different sources can be brought together into a single topic map and integrated into a meaningful whole.
- Different web sites can be integrated with one another since they will know what on one web site corresponds to what on the other.

This list is hardly comprehensive, since the thinking about how to exploit these possibilities is still quite young.

Nothing like this facility exists in traditional subject-based classification, though XFML provides the same facility. However, although XFML does make it possible to tell when two terms from different XFML maps represent the same concept there is only limited help in that. If the different maps give the same concept different names, or different parents, or place them in difference facets, then XFML cannot represent the resulting merged concept. To some extent this is a problem inherent in how hierarchical classification systems model the world, which topic maps do not share.

6. Conclusion

What this paper has tried to show is that topic maps provide a common reference model that can be used to explain how to understand many common techniques from library science and information architecture. It has also showed

Rendered by www.RenderX.com

how these techniques can be implemented using topic maps, and how topic maps can go far beyond the possibilities provided by traditional techniques.

By using topic maps to represent metadata and subject-based classification it is possible to reuse existing classifications and classification techniques, while at the same time describing the world more precisely where desired.

Acknowledgements

Many thanks to Marte Brenne for references to useful definitions of some of the library science terms used in this paper.

Many thanks to Murray Altheim for expanding my view of what faceted classification is and for providing useful references.

Thanks also to Isabella Kubosch, Magnus Halle, Runar Eggen, and Tord Høivik for useful references.

Many thanks to Pam Gennusa, Sylvia Schwab and Steve Pepper for reading and commenting on early versions of the paper.

Bibliography

[Ahmed] Topic Map Design Patterns For Information Architecture, Kal Ahmed, XML 2003, December 2003. http://www.techquila.com/tmsinia.html

[van Dijck] -Introduction to XFML, Peter van Dijck, XML.com, January 22, 2003. http://www.xml.com/pub/a/2003/01/22/xfml.html

[Dublin Core] RFC 2413: Dublin Core Metadata for Resource Discovery, Weibel, S.; Kunze, J.; Lagoze, C.; Wolf, M., The Internet Society, September 1998. http://www.ietf.org/rfc/rfc2413.txt

[Garshol] XFML Ontology and Converter, Lars Marius Garshol, Ontopia, forthcoming. http://psi.ontopia.net/xfml/

[ISO2788] ISO 2788:1986 — Guidelines for the establishment and development of monolingual thesauri, International Organization for Standardization (ISO), 1986.

[ISO5964] ISO 5964:1985 — Guidelines for the establishment and development of multilingual thesauri, International Organization for Standardization (ISO), 1985.

[Pepper] The TAO of Topic Maps, Steve Pepper, Ontopia, April 2002. http://www.ontopia.net/topicmaps/materials/tao.html

[Ranganathan] Ranganathan for IAs, Mike Steckel, Boxes and Arrows (online magazine), October 7, 2002. http://www.boxesandarrows.com/archives/ranganathan_for_ias.php

[Svenonius] The Intellectual Foundation of Information Organization, Elaine Svenonius, MIT Press, 2000, ISBN 0-262-19433-3.

[The XML Papers] The XML Papers, Garshol, L. M.; Pepper, S., XML 2002 Conference Proceedings, IDEAlliance, December 2002. http://www.ontopia.net/topicmaps/materials/xmlconf.html

Glossary

TMCL	Topic Map Constraint Language

Rendered by www.RenderX.com

XFML eXchangeable Faceted Metadata Language

Biography

Lars Marius **Garshol**
Development Manager
Ontopia
Oslo
Norway
larsga@ontopia.net

Lars Marius Garshol is currently Development Manager at Ontopia, a leading topic map software vendor. He has been active in the XML and topic map communities as a speaker, consultant, open source developer, and technology creator for a number of years. He helped develop the standard SAX API for development and translated it to Python, and wrote an open-source validating XML parser in Python.

Lars Marius has also been responsible for adding Unicode support to the Opera web browser. His book on Definitive XML Application Development, was published by Prentice-Hall in its Charles Goldfarb series. Lars Marius is one of the editors of the ISO Topic Map Query Language standard, and also co-editor of the Topic Map data model.

Rendered by www.RenderX.com

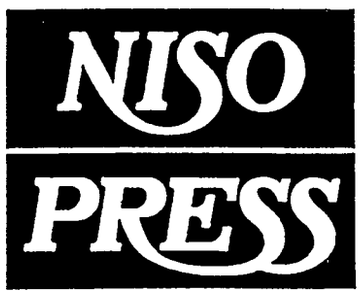
ANSI/NISO Z39.19-1993
(Revision of Z39.19-1980)

ISSN: 1041-5653

Guidelines for the Construction, Format, and Management of Monolingual Thesauri

Abstract: A thesaurus is a controlled vocabulary arranged in a known order and structured so that equivalence, homographic, hierarchical, and associative relationships among terms are displayed clearly and identified by standardized relationship indicators that are employed reciprocally. The primary purposes of a thesaurus are (a) to facilitate retrieval of documents and (b) to achieve consistency in the indexing of written or otherwise recorded documents and other items, mainly for postcoordinate information storage and retrieval systems. This standard provides guidelines for constructing monolingual thesauri: formulating the descriptors, establishing relationships among terms, and effectively presenting the information in print and on a screen. It also includes thesaurus maintenance procedures and recommended features of thesaurus management systems.

An American National Standard
Developed by the
National Information Standards Organization
Approved August 30, 1993 by the
American National Standards Institute



Bethesda, Maryland, U.S.A.

Published by
NISO Press
P.O. Box 1056
Bethesda, MD 20827

Copyright ©1994 by the National Information Standards Organization
All rights reserved under International and Pan-American Copyright Conventions. No part of this book may be reproduced or transmitted in any form or by any means, electronic or mechanical, including photocopy, recording, or any information storage or retrieval system, without prior permission in writing from the publisher. All inquiries should be addressed to NISO Press, P.O. Box 1056, Bethesda, MD 20827.

ISSN: 1041-5653 National Information Standards Series
ISBN: 1-880124-04-1

Printed in the United States of America

This paper meets the requirements of ANSI/NISO Z39.48-1992 (Permanence of Paper).

Library of Congress Cataloging-in-Publication Data

National Information Standards Organization (U.S.)
Guidelines for the construction, format, and management of monolingual thesauri / developed by the National Information Standards Organization : approved August 30, 1993, by the American National Standards Institute.
p. cm. — (National information standards series, ISSN 1041-5653; ANSI/NISO Z39.19-1993)
Rev. ed. of: American national standard guidelines for thesaurus structure, construction, and use. New York : ANSI, c1980.
Includes bibliographical references (p.) and index.
ISBN 1-880124-04-1 : $49.00
1. Subject headings—Standards—United States. I. American National Standards Institute. II. American National Standards Institute. Subcommittee 25 on Thesaurus Rules and Conventions. American national standard guidelines for thesaurus structure, construction, and use. III. Title. IV. Series.
Z695.N36 1994 93-42543
025.4'9—dc20 CIP

Contents

Foreword viii
Abbreviations, Codes, and Conventions Used in This Standard xii
1. Introduction 1
1.1. Purpose and Structure 1
1.1.1. Purpose 1
1.1.2. Structure 1
1.2. Vocabulary Control 1

2. Scope of the Standard 1
2.1. Applications of the Standard 2
2.2. Monolingual Thesauri 2
2.3. Types of Terms 2

3. Scope, Form, and Choice of Descriptors 2
3.1. Single-Word vs. Multiword Descriptors 2
3.2. Scope of Descriptors 2
3.2.1. Parenthetical Qualifiers 2
3.2.2. Scope Notes 3
3.2.2.1. Reciprocal Scope Notes 3
3.3. Types of Concept 3
3.3.1. Unique Entities 4
3.4. Grammatical Forms of Descriptors 4
3.4.1. Nouns and Noun Phrases 4
3.4.1.1. Verbal Nouns 4
3.4.1.2. Noun Phrases 4
3.4.1.2.1. Adjectival Noun Phrases 4
3.4.1.2.2. Prepositional Noun Phrases 5
3.4.2. Adjectives 5
3.4.2.1. Limiting the Number of Compound Descriptors 5
3.4.2.2. Economy of Cross-References 5
3.4.3. Adverbs 5
3.4.4. Initial Articles 5
3.4.4.1. Deletion 5
3.4.4.2. Retention 5
3.5. Singular and Plural Forms 6
3.5.1. Count Nouns 6
3.5.1.1. Exception to Plural Count Nouns 6
3.5.2. Noncount (Mass) Nouns 6
3.5.2.1. Abstract Concepts 6
3.5.2.2. Unique Entities 6
3.5.3. Coexistence of Singular and Plural Forms 6
3.6. Selection of Preferred Form 7
3.6.1. Usage 7
3.6.1.1. Literary Warrant 7
3.6.1.2. Published Authorities 7
3.6.1.3. Other Indexing Languages 7
3.6.1.4. Personal Authorities 7
3.6.2. Spelling 7
3.6.2.1. Literary Warrant 7

(continued)

(3. Scope, Form, and Choice of Descriptors, continued)
3.6.2.2. Authorities 7
3.6.3. Abbreviations, Initialisms, and Acronyms 7
3.6.3.1. Preference for Abbreviation 7
3.6.3.2. Preference for Full Form 7
3.6.3.2.1. General Use 7
3.6.3.2.2. Ambiguity 7
3.6.4. Neologisms, Slang, and Jargon 8
3.6.4.1. Provisional Terms 8
3.6.4.2. Cross-References from Slang Terms 8
3.6.5. Common Names and Trade Names 8
3.6.5.1. Preference for Common Name 8
3.6.5.2. Trademarks 8
3.6.5.3. Loss of Protected Status 8
3.6.6. Popular and Scientific Names 8
3.6.7. Loanwords, Translations of Loanwords, and Foreign-Language Equivalents 8
3.6.7.1. Loanwords 8
3.6.7.2. Translations of Loanwords 8
3.6.7.3. Foreign-Language Equivalents 9
3.6.8. Proper Names 9
3.6.8.1. Place Names 9
3.7. Capitalization, Punctuation, and Nonalphabetic Characters 9
3.7.1. Capitalization 9
3.7.2. Nonalphabetic Characters 10
3.7.2.1. Parentheses 10
3.7.2.2. Hyphens 10
3.7.2.3. Apostrophes 10
3.7.2.3.1. Possessive Case 10
3.7.2.3.2. Proper Names 10
3.7.2.4. Diacritical Marks 10
3.7.2.5. Other Symbols and Punctuation Marks 10
3.7.3. Romanization 10

4. Compound Terms 10
4.1. General 10
4.1.1. Purpose of Guidelines on Compound Terms 11
4.1.2. Precoordination and Postcoordination 11
4.1.3. Factors to Be Considered 11
4.1.4. Elements of Compound Terms 11
4.2. Criteria for Establishing Compound Terms as Descriptors 12
4.3. Criteria for Determining when Compound Terms Should Be Split 12
4.3.1. Factors to Be Considered 12
4.3.2. Hierarchical Structure 12
4.3.2.1. Focus and Difference 12
4.3.2.2. Transitive Action 12
4.3.2.3. Intransitive Action 13
4.4. Compound Descriptors vs. Node Labels 13
4.5. Order of Words in Compound Terms 13
4.5.1. Cross-References from Inversions 13

5. Relationships 13
5.1. General 13
5.1.1. Semantic Linking 13

(continued)

(5. Relationships, continued)
5.1.2. Kinds of Relationship 13
5.2. The Equivalence Relationship 15
5.2.1. General 15
5.2.2. Synonyms 15
5.2.3. Lexical Variants 16
5.2.4. Quasi-Synonyms 16
5.2.4.1. Generic Posting 16
5.2.5. Cross-References to Elements of Compound Terms 16
5.3. The Hierarchical Relationship 16
5.3.1. The Generic Relationship 17
5.3.1.1. Codes for the Generic Relationship 17
5.3.2. The Whole-Part Relationship 17
5.3.2.1. Codes for the Whole-Part Relationship 18
5.3.2.2. Parts of Multiple Wholes 18
5.3.3. The Instance Relationship 18
5.3.3.1. Codes for the Instance Relationship 18
5.3.4. Polyhierarchical Relationships 18
5.3.5. Node Labels in Hierarchies 18
5.4. The Associative Relationship 19
5.4.1. Relationships Between Descriptors Belonging to the Same Hierarchy 19
5.4.1.1. Relationships Between Overlapping Sibling Terms 19
5.4.1.2. Relationships Between Mutually Exclusive Sibling Terms 19
5.4.1.3. Derivational Relationships 19
5.4.2. Relationships Between Descriptors Belonging to Different Hierarchies 20
5.4.3. Node Labels for Related Terms 20
5.4.4. Specifying Types of Related Term References 20

6. Print Display 21
6.1. The Equivalence Relationship 21
6.1.1. Inverted USE References and Permuted Lists 21
6.1.2. Inverted USE References and Hierarchical Display 21
6.1.3. Juxtaposition of Terms 21
6.2. Hierarchical Relationships 22
6.2.1. Conventional Flat Thesaurus Structure 22
6.2.2. Multilevel Hierarchy Within Alphabetic Display 22
6.2.2.1. Multilevel Broader and Narrower Terms 22
6.2.2.2. Generic Structure 22
6.2.3. Tree Structure 22
6.2.3.1. Notation 23
6.2.4. Graphic Displays 23
6.2.5. The Top Term Structure 23
6.2.6. Two-Way Hierarchical Structure 23
6.2.7. Broad Categories 23
6.3. General Format Considerations 23
6.3.1. Components of a Printed Thesaurus 23
6.3.2. Introduction 23
6.3.3. Typography 24
6.3.4. Capitals/Lowercase Letters 24
6.3.5. Filing 24
6.3.5.1. Alphabetic Characters 24
6.3.5.2. Numerals 24
6.3.5.3. Non-Alphanumeric Characters 24
(continued)

(6. Print Display, continued)
6.3.6. Running Heads 25
6.4. Multiple Versions of Printed Thesauri 25

7. Screen Display 25
7.1. User Categories 25
7.2. Factors to Be Considered 25
7.3. Types of Display 26
7.3.1. Alphabetical Listing 26
7.3.2. Permuted Displays 26
7.3.3. Hierarchies and Classifications 26
7.3.4. Term Detail 26
7.3.5. Graphic Displays 26
7.4. Documentation 26
7.5. Relation of Thesaurus Character Set to Search Commands 26

8. Thesaurus Construction 27
8.1. Avoidance of Duplicate Work 27
8.2. Determination of Structure and Display Format 27
8.3. Methods 27
8.3.1. The Committee Approach 27
8.3.2. The Empirical Approach 27
8.3.3. Combination of Methods 27
8.4. Machine Assistance 27
8.5. Term Records 28
8.6. Term Verification 28
8.6.1. Candidate Terms 28
8.6.2. Provisional Terms 28
8.7. Level of Specificity 28
8.8. Unassigned Descriptors 29
8.9. Announcement and Deposit of Published Thesauri 29
8.9.1. Notification of Intent 29
8.9.2. Pilot Thesaurus 29
8.9.3. Deposit with a Clearinghouse 29

9. Thesaurus Maintenance 29
9.1. Updating the Vocabulary 29
9.1.1. Addition of Descriptors 29
9.1.2. Modification of Existing Descriptors 29
9.1.3. Deletion of Terms 30
9.2. Vocabulary Updates and Database Records 30
9.3. History of Changes 30

10. Thesaurus Management Systems 31
10.1. Typography and Sorting 31
10.2. Term Records and Display Format 31
10.3. System Constraints 31
10.4. Searching 31
10.5. Editing 31
10.6. Error Checking 31
10.7. Cross-References 32
10.8. Term Deletion 32
10.9. Candidate Terms 32

(continued)

(10. Thesaurus Management Systems, continued)
10.10. Field Definition 32
10.11. Reports 33
10.12. Subsets 33
10.12.1. Copying Function 33
10.12.2. Microthesauri 33

References 34

Glossary of Terms 35

Appendixes
A. Supplementary Figures 39
B. Minithesaurus of Thesaurus Terms 61
C. User Interfaces 62
D. Thesaurus Clearinghouses 63
E. Bibliography of Manuals on Thesaurus Construction 64

Index 65

List of Figures

Figure 1. Semantic Relationships in a Thesaurus 14

Appendix A: Supplementary Figures

A1. The Relationship Indicators USE AND and USED FOR+ 39
A2. KWIC Index to a Thesaurus 40
A3. KWOC Index to a Thesaurus 41
A4. Flat Thesaurus Structure 42
A5. Multilevel Broader and Narrower Terms 43
A6. Generic Structure 44
A7. Tree Structure 45
A8. Hierarchical Thesaurus Display Featuring Node Labels and Line Numbers 46
A9. Graphic Display 47
A10. Top Terms in an Alphabetic Thesaurus 48
A11. Hierarchical Display of Top Terms 49
A12. Two-Way Hierarchical Display 50
A13. Category List 51
A14. Printed Alphabetic Thesaurus Designed for End-Users 52
A15. Printed Alphabetic Thesaurus Designed for Indexers and Searchers 53
A16. KWIC Screen Display of Thesaurus Terms 54
A17. KWOC Screen Display of Thesaurus Terms 55
A18. Screen Display of Term Record 56
A19. Thesaurus Term Review Form 57
A20. Menu of a Thesaurus Management System 58
A21. Screens for Modification and Deletion of Terms 59
A22. Screen Display of Term Record Featuring Specialized Codes in a Company Thesaurus 60

Foreword

(This foreword is not part of the American National Standard Guidelines for the Construction, Format, and Management of Monolingual Thesauri, ANSI/NISO Z39.19-1993. It is included for information only.)

The first edition of this standard, published in 1974, was prepared by Subcommittee 25 on Thesaurus Rules and Conventions of American National Standards Committee Z39 on Standardization in the Field of Library Work, Documentation, and Related Publishing Practices (later known as Library and Information Sciences and Related Publishing Practices) [1]. The subcommittee drew heavily on standards of practice developed by the Engineers Joint Council, the Committee on Scientific and Technical Information of the Federal Council for Science and Technology, and UNESCO [2,3,4]. In 1980 ANSI published a revised edition of the thesaurus standard. In 1985, the National Information Standards Organization, the successor of American National Standards Committee Z39, balloted to reaffirm the standard. Two "no" votes led to a default ballot in 1987, which resulted in additional recommendations for revision.

In 1988 Standards Committee PP was appointed to draft a revised standard. The committee members, chosen for their expertise in both the theory and the practice of the development and use of thesauri in information systems, represented a wide range of subject interests (aerospace, business, medicine, social sciences, information science, and the arts).

The current revision borrows heavily from the corresponding international and British standards [5,6]. It differs substantially in structure and wording from the 1980 American edition, but supports most of its recommendations. An important change is in the handling of the whole-part relationship, which is now treated as a special case of the hierarchical relationship rather than as an associative relationship.

Another notable change is the deletion of the word *use* from the title because this standard does not prescribe how thesauri are to be used. The phrase *monolingual thesauri* has been added to the title to distinguish this standard from a separate standard for multilingual thesauri [7].

This edition continues to emphasize the role of thesauri in information storage and retrieval systems, but recognizes their applicability to other fields, such as knowledge engineering. New sections deal with screen display and thesaurus management systems.

Unlike the previous edition, the revised standard does not take a monolithic approach to thesaurus structure and display, but instead points out the advantages and disadvantages of various formats. Many of these formats are illustrated in an appendix of supplementary figures.

Suggestions for improving this standard are welcome. They should be sent to the National Information Standards Organization, P.O. Box 1056, Bethesda, MD 20827, telephone (301) 975-2814.

At the time it approved this standard, NISO had the following voting members:

NISO Voting Members

American Association of Law Libraries
Gary J. Bravy

American Chemical Society
Robert S. Tannehill, Jr.
Leon R. Blauvelt (Alt)

American Library Association
Myron Chace
Glenn Patton (Alt)

American Psychological Association
Maurine F. Jackson

American Society for Information Science
Nolan Pope

American Society of Indexers
Patricia S. Kuhr

American Theological Library Association
Myron Chace

Apple Computer, Inc.
Karen Higginbottom

Art Libraries Society of North America
Patricia J. Barnett
Pamela J. Parry (Alt)

Association of American University Presses
Mary Lou Menches

Association of Information and Dissemination Centers
Bruce H. Kiesel

Association for Information and Image Management
Marilyn Courtot

Association of Jewish Libraries
Bella Hass Weinberg
Pearl Berger (Alt)

The Association for Recorded Sound Collections
Donald McCormick
Martin Silver (Alt)

Association of Research Libraries
Duane Webster

AT&T Bell Labs
M. E. Brennan

Baker & Taylor Books
Christian K. Larew
Stephanie Lanzalotto (Alt)

The Boeing Company
Steven C. Hill
Michael Crandall (Alt)

Book Manufacturers' Institute
Stephen P. Snyder

CLSI, Inc.
Andy Lukes

Colorado Alliance of Research Libraries
Ward Shaw

Council of National Library and Information Associations
Thomas J. Kemp

Data Research Associates, Inc.
Michael J. Mellinger
James Michael (Alt)

Digital Equipment Corporation
Joan Blair

Dynix
Rick Wilson

EBSCO Industries, Inc.
Mary Beth Vanderpoorten

Engineering Information, Inc.
Eric Johnson
Mary Berger (Alt)

The Faxon Co., Inc.
Fritz Schwartz
Joe Santosuosso (Alt)

Gaylord Information Systems
Robert Riley
Bradley McLean (Alt)

IBM Corporation
Peggy Federhart

Indiana Cooperative Library Services Authority
Barbara Evans Markuson
Janice Cox (Alt)

Library Binding Institute
Sally Grauer

Library of Congress
Winston Tabb
Sally H. McCallum (Alt)

Mead Data Central
Peter Ryall
Dave Withers (Alt)

Medical Library Association
Rick B. Forsman
Raymond A. Palmer (Alt)

MINITEX
Anita Anker Branin
William DeJohn (Alt)

Music Library Association
Lenore Coral
Geraldine Ostrove (Alt)

National Agricultural Library
Joseph H. Howard
Gary K. McCone (Alt)

National Archives and Records Administration
Alan Calmes

National Federation of Abstracting and Information Services
Sarah Syen

National Institute of Standards and Technology, Office of Information Services
Jeff Harrison
Marietta Nelson (Alt)

National Library of Medicine
Lois Ann Colaianni

OCLC, Inc.
Kate Nevins
Don Muccino (Alt)

OHIONET
Greg Pronevitz

Optical Publishing Association
John Nairn
Richard A. Bowers (Alt)

PALINET
James E. Rush

Pittsburgh Regional Library Center
Mary Lynn Kingston

Readmore Academic Services
Sandra J. Gurshman
Dan Tonkery (Alt)

The Research Libraries Group, Inc.
Wayne Davison
Kathy Bales (Alt)

Society of American Archivists
Christine Ward
Victoria Irons Walch (Alt)

Software AG of North America, Inc.
James J. Kopp
James E. Emerson (Alt)

Special Libraries Association
Audrey N. Grosch

SUNY/OCLC Network
Glyn T. Evans
David Forsythe (Alt)

UMI
Don Willis
John Brooks (Alt)

Unisys Corporation
Bill Payne

U.S. Department of Commerce, Printing and Publishing Division
William S. Lofquist

U.S. Department of Defense, Defense Technical Information Center
Gretchen Schlag
Claire Tozier (Alt)

U.S. Department of Defense, Joint Medical Library
Diane Zehnpfennig

U.S. Department of Energy, Office of Scientific and Technical Information
Mary Hall
Nancy Hardin (Alt)

U.S. ISBN Maintenance Agency
Emery Koltay

U.S. National Commission on Libraries and Information Science
Peter Young
Sandra N. Milevski (Alt)

VTLS
Vinod Chachra

H.W. Wilson Company
George I. Lewicky
Ann Case (Alt)

Standards Committee PP, which developed this standard, had the following members:

Bella Hass Weinberg, Chair
St. John's University

Ronald L. Buchan
NASA Scientific and Technical Information Program

Lois Lyman
Digital Equipment Corporation

Toni Petersen
Art and Architecture Thesaurus

Caroline Reyes
NewsBank/Readex

Peri Schuyler
National Library of Medicine

Hans H. Wellisch
Emeritus, University of Maryland

NISO Board of Directors

At the time NISO approved this standard, the following individuals served on its Board of Directors:

James E. Rush, Chairperson
PALINET

Michael J. Mellinger, Vice Chair/Chair-elect
Data Research Associates

Paul Evan Peters, Immediate Past Chairperson
Coalition for Networked Information

Heike Kordish, Treasurer
New York Public Library

Patricia R. Harris, Executive Director
National Information Standards Organization

Directors Representing Libraries

Susan Vita
Library of Congress

Shirley Kistler Baker
Washington University

Nolan Pope
University of Wisconsin-Madison

Directors Representing Information Services

Michael J. McGill
University of Michigan

Wilhelm Bartenbach
Engineering Information

Bonnie Lawlor
Institute for Scientific Information, Inc.

Directors Representing Publishing

Joel H. Baron
The Faxon Company, Inc.

Constance U. Greaser
American Honda

Marjorie Hlava
Access Innovations, Inc.

References

1. American National Standard Guidelines for Thesaurus Structure, Construction, and Use. New York: American National Standards Institute, 1974. (ANSI Z39.19-1974).

2. Thesaurus Rules and Conventions. New York: Engineers Joint Council, November 1966.

3. Guidelines for the Development of Information Retrieval Thesauri. U.S. Federal Council for Science and Technology, Committee on Scientific and Technical Information (COSATI). Washington, D.C.: Government Printing Office, 1967.

4. Guidelines for the Establishment and Development of Monolingual Scientific and Technical Thesauri for Information Retrieval. Paris: United Nations Educational, Scientific and Cultural Organization, July 1970.

5. Documentation—Guidelines for the Establishment and Development of Monolingual Thesauri, 2nd ed. Geneva: International Organization for Standardization, 1986. (ISO 2788-1986).

6. Guidelines for the Establishment and Development of Monolingual Thesauri. London: British Standards Institution, 1979. (BS 5723:1979). Superseded by: British Standard Guide to Establishment and Development of Monolingual Thesauri. (BS 5723:1987=ISO 2788:1986).

7. Documentation—Guidelines for the Establishment and Development of Multilingual Thesauri. Geneva: International Organization for Standardization, 1985. (ISO 5964-1985).

Abbreviations (Thesaurus Codes) and Conventions Used in This Standard

Abbreviations (Thesaurus Codes) and Relationship Indicators:

BT = broader term
BTG = broader term (generic)
BTI = broader term (instance)
BTP = broader term (partitive)
GS = generic structure
HN = history note
NT = narrower term
NTG = narrower term (generic)
NTI = narrower term (instance)
NTP = narrower term (partitive)
RT = related term
SEE = equivalent to U (USE)
SN = scope note
TT = top term
U = use
UF = used for
UF+ = used for . . . and . . .
USE+ = use . . . and . . .
X = see-from (equivalent to UF); reciprocal of *see*

Conventions:

Descriptors are in **boldface**.

Words discussed in the text are enclosed in 'single quotation marks'.

Underlined terms are defined in the glossary.

Guidelines for the Construction, Format, and Management of Monolingual Thesauri

1. Introduction

A thesaurus*, for the purposes of this standard, is a controlled vocabulary of terms in natural language that are designed for postcoordination. The need to control the formation and use of terms stems mainly from two basic features of natural language, namely synonyms (different terms representing the same concept) and polysemes or homographs (terms with the same spelling representing different concepts). The controlled vocabulary is established by information specialists or lexicographers and is generally employed in indexing.

Indexing is the process by which subject terms or classification symbols are assigned to concepts dealt with in documents. It includes any system in which the selection and organization of indexing terms call for human intellectual decisions, regardless of whether computer assistance is also used to store and manipulate these terms or to identify documents to which certain terms or combinations of terms have been assigned. The effectiveness of indexing as a means for identifying and retrieving documents depends upon a well-constructed indexing language.

Research in the field of information science has shown that controlled vocabularies improve both precision and recall in searching. For example, they improve precision by defining the scope of terms; they increase recall by retrieving documents that employ different terms for the same concept. (Some of the manuals of thesaurus construction listed in Appendix E cite studies that have demonstrated these advantages of controlled vocabularies.)

Thesauri may be used in computer-assisted indexing, but this standard is primarily oriented toward thesauri employed for assignment indexing by humans (as opposed to derivative indexing by machines).

It is important to note, however, that thesauri are not limited to systems that apply vocabulary control at the indexing stage. Thesauri may also be helpful to the user searching unindexed, i.e., natural language, databases. The value of thesauri is currently recognized as well by researchers in knowledge engineering and hypertext. The many uses to which thesauri may be put are recognized in the standard, but not explicitly discussed.

* Underlined terms are defined in the glossary.

1.1. Purpose and Structure

A thesaurus may be considered in terms of both its purpose and its structure.

1.1.1. Purpose

Four principal purposes are served by a thesaurus:

a) *Translation.* To provide a means for translating the natural language of authors, indexers, and users into a controlled vocabulary used for indexing and retrieval.
b) *Consistency.* To promote consistency in the assignment of index terms.
c) *Indication of Relationships.* To indicate semantic relationships among terms.
d) *Retrieval.* To serve as a searching aid in retrieval of documents.

1.1.2. Structure

A thesaurus displays through its structure the equivalence (synonymous), hierarchical, and associative relationships among terms. These relationships are defined in section 5. Print and screen display formats used to show these relationships are recommended in sections 6 and 7, respectively.

1.2 . Vocabulary Control

Vocabulary control in a thesaurus is achieved through three principal means: (a) the delineation of the scope, or meaning, of descriptors (see section 3.2); (b) the linking of synonymous and nearly synonymous terms through the equivalence relationship (see section 5.2); and (c) the disambiguation of homographs (see section 3.2.1).

2. Scope of the Standard

This standard presents guidelines and conventions for the contents, display, methods of construction, and maintenance of a monolingual thesaurus. It also notes the recommended features of a thesaurus management system.

The standard deals with some aspects of term selection, in that it contains recommended procedures for vocabulary control, but it is particularly concerned with means for establishing and displaying certain kinds of relationships between terms.

This standard should be regarded as a set of recommendations based on preferred techniques

and procedures. Optional procedures are, however, sometimes described, e.g., for the display of inter-term relationships in a thesaurus.

The conventions used in this standard to indicate the force of recommendations are: *shall* (required for meeting the standard), *should* (recommended), and *may* (optional).

2.1. Applications of the Standard

The standard assumes that thesauri will be applied mainly in postcoordinate information storage and retrieval systems. Guidelines for the development of lists of subheadings designed for precoordination with descriptors are therefore not included in this standard, but this does not preclude the use of thesauri in precoordinate indexing systems.

The standard also does not provide guidelines for human or computer-aided indexing based on thesauri. For indexing procedures and practices, see ANSI Z39.4.

2.2. Monolingual Thesauri

The recommendations contained in this standard relate specifically to *monolingual* thesauri, with particular reference to English. No consideration has been given to the special requirements of other languages, nor to those of multilingual thesauri in which terms are listed in parallel sequences showing conceptual equivalences in more than one language.

2.3. Types of Terms

This standard deals only with the formulation, organization, and display of terms that form a controlled subset of natural language. It does not suggest procedures for organizing and displaying headings consisting of mathematical and chemical formulas, nor for establishing proper names. (For information on these, refer to appropriate standards, such as *Anglo-American Cataloguing Rules*, 2nd ed. [AACR2].) Such headings may be included in thesauri, however, and relationships among them may be indicated through the devices presented in this standard.

3. Scope, Form, and Choice of Descriptors

This section deals with the principles of descriptor selection and the determination of the correct form. Literary warrant is the guiding principle for the selection of the preferred form of a descriptor. Section 3.2.1 treats the form of cross-references from nonpreferred terms to preferred terms. In this section, the phrase 'rather than' is used to contrast preferred and nonpreferred terms. Where two synonyms are juxtaposed without indication of a preference, they are separated by a slash.

The development of an initial set of descriptors is dealt with in section 8.

3.1. Single-Word vs. Multiword Descriptors

Each descriptor included in a thesaurus should represent a single concept (or unit of thought). A concept may be expressed by a single-word term or by a multiword term. Because of the difficulty of defining 'single concept' in the latter case, objective criteria for dealing with compound terms are provided in section 4.

3.2. Scope of Descriptors

The scope of descriptors is restricted to selected meanings within the domain of the thesaurus. Each descriptor should be formulated in such a way that it conveys the intended scope to any user of the thesaurus. Terms whose meanings overlap in general usage, and homographs, i.e., terms with identical spellings but different meanings, should be avoided as far as possible in the selection of descriptors. Literary warrant, i.e., the occurrence of terms in documents, and other criteria may, however, justify the use of such terms as descriptors. Section 3.2.1 deals with methods of disambiguating homographs. Section 3.2.2 explains the use of scope notes for terms whose meanings overlap.

3.2.1. Parenthetical Qualifiers

The use of homographic terms as descriptors requires clarification of their meaning through a qualifier ('gloss' in linguistic terminology). The qualifier, which is enclosed in parentheses, is part of the descriptor. The qualifier itself may be a descriptor, often a broader term than the one being qualified. It should be as brief as possible, ideally consisting of one word, but should not be a homograph. Qualifiers should be standardized within a given thesaurus to the extent possible, e.g., 'biology' and 'bioscience' should not both be used as qualifiers.

A qualifier is not a scope note; however, a qualified descriptor may have a scope note appended to it. For guidelines on scope notes, see section 3.2.2.

Qualifiers should also be added to entry terms when their meaning is ambiguous.

Examples:

cranes (birds)
cranes (lifting equipment)

mercury (metal)
Mercury (planet)
Mercury (Roman deity)

seals (animals)
seals (law)
seals (numismatics)

socialization (economics)
socialization (social psychology)

a) A qualifier should be added to each homographic term, even when one is used in the primary sense of the domain and the second in a different sense. For example, **cranes (lifting equipment)** should be the descriptor in an engineering thesaurus that also includes **cranes (birds)**.
b) A parenthetical qualifier should be appended to homographic terms even when a descriptor is used in only one of its meanings within a thesaurus, e.g., **shells (structures)** in an engineering thesaurus. This facilitates cross-database searching and mapping of descriptors in disparate domains.
Examples:
developing (photography)
organism (philosophy)
translation (genetics)
c) Parenthetical qualifiers should not be used to represent compound concepts, e.g., 'cookery (zucchini)' or 'pipes (plastic).' 'Plastic' is used in the latter example to indicate a type of pipe rather than to disambiguate the word 'pipe.' Appropriate uses of qualifiers with the term 'pipes' would be: **pipes (musical instruments)** and **pipes (smoking implements)**. (See section 4 for guidelines on formulating compound descriptors.)
d) Selection of a compound term as a descriptor is preferred to use of a parenthetical qualifier with a single-word term, if usage permits, i.e., if the compound occurs in natural language.
Examples:
phonograph records *rather than* records (phonograph)
religious tolerance *rather than* tolerance (religion)

Both of these compound terms are natural expressions that may be selected for thesauri in the fields of music and religion, respectively.

3.2.2. Scope Notes

A scope note is used to restrict or expand the application of a descriptor, to distinguish between descriptors that have overlapping meanings in natural language, or to provide other advice on term usage to either the indexer or the searcher. A scope note should state the chosen meaning of a descriptor; it may also indicate other meanings that are recognized in natural language, but which have been deliberately excluded from the controlled vocabulary. A scope note, unlike a parenthetical qualifier, is not part of a descriptor. While qualifiers are generally added only to homographs, a scope note (SN) may be supplied for any descriptor.

Example:
illuminations
SN Includes both the ornamental decoration and the illustrations in manuscripts as well as in some early printed books, if done by hand.

The scope of descriptors is indicated additionally through the semantic relationships represented in the thesaurus (see section 5.2). Changes in the scope of descriptors should be recorded in history notes (see section 9.3).

3.2.2.1. Reciprocal Scope Notes

When reference is made to other descriptors in a scope note, a reciprocal scope note should generally be provided for each descriptor mentioned.

Examples:
dogtrots
SN Passages sharing a roof common with the rest of a building, connecting two parts of a log house of the American folk art tradition. Distinguished from **breezeways** by its folk architecture tradition and log house context and its common roof.
breezeways
SN Roofed passages connecting two parts of a house or a house and a garage; common after 1930. Distinct from **dogtrots**, which occur in folk architecture log houses.

Even where the scope of only one of the descriptors requires clarification, it is useful to note in the term record for the second descriptor that it has been cited in a scope note of a different descriptor, e.g.:

information science
X SN **library science**

The *X* indicates that there is a reference from the scope note of **library science** to **information science**. This reciprocal reference will ensure that when a change is made to one of the descriptors, or it is deleted, the effect on the other descriptor will be considered.

3.3. Types of Concept

The concepts represented by descriptors can be grouped into general types. The compiler of a

thesaurus needs to be aware of these types because the type of concept may affect some of the procedures used in thesaurus construction; for example, the choice of singular or plural form (see section 3.5) and applying a test for the validity of a hierarchy (see section 5.3.1). The listed types are not exhaustive.

a) things and their physical parts
   Examples:
   birds
   carburetors
   microforms
   mountains
   oil paintings
   teddy bears

b) materials
   Examples:
   adhesives
   mustard gas
   oxygen
   paints
   water

c) activities or processes
   Examples:
   acidification
   miniature golf
   painting
   parenting
   scuba diving
   sewing

d) events or occurrences
   Examples:
   birthdays
   civil wars
   holidays
   revolutions

e) properties or states of persons, things, materials, or actions
   Examples:
   consciousness
   elasticity
   personality
   speed
   texture

f) disciplines or subject fields
   Examples:
   anthropology
   information science
   organic chemistry
   theology

g) units of measurement
   Examples:
   hertz
   kilometers

3.3.1. Unique Entities

Unique entities, or "classes-of-one," are usually expressed as proper nouns.

Examples:
Aristotle
Earth
Fourth of July
Library of Congress
Magna Carta
World Health Organization
Zimbabwe

3.4. Grammatical Forms of Descriptors

The following sections indicate the preferred grammatical forms of descriptors. Entry terms from nonpreferred grammatical forms may be provided.

3.4.1. Nouns and Noun Phrases

The grammatical form of a descriptor should be a noun or noun phrase. For an exception to this recommendation, see section 3.4.2.

Examples:
drawings
liver
space shuttles

3.4.1.1. Verbal Nouns

Verbs expressed as infinitives (without 'to') or participles should not be used alone as descriptors. Activities should be represented by nouns or gerunds.

Examples:

catalysis	*rather than*	catalyze
cookery	*rather than*	cook
freezing	*rather than*	frozen
reading	*rather than*	read
distillation	*rather than*	distill

3.4.1.2. Noun Phrases

Noun phrases are compound terms that may be established as descriptors if they represent a single concept (see section 4). Noun phrases occur in two forms: premodified, or adjectival; and postmodified, or prepositional.

3.4.1.2.1. Adjectival Noun Phrases

Adjectival noun phrases, which are generally premodified, are the preferred form.

Examples:
basal metabolism
cold fusion
historical drama
public television
rapid transit

3.4.1.2.2. Prepositional Noun Phrases

Prepositional noun phrases are generally postmodified. When possible, noun phrases should exclude prepositions; for example, use 'carbohydrate metabolism' rather than 'metabolism of carbohydrates,' and 'children's hospitals' rather than 'hospitals for children.'

Descriptors in the form of prepositional noun phrases should be restricted to concepts that cannot be expressed in any other way, or that have become idiomatic.

Examples:

accessories after the fact
burden of proof
cream of tartar
coats of arms
plaster of Paris
prisoners of war
sergeants-at-arms
strength of materials

3.4.2. Adjectives

Adjectives and adjectival phrases used alone may be established as descriptors in thesauri in the special circumstances considered below.

3.4.2.1. Limiting the Number of Compound Descriptors

As an alternative to the creation of multiple compound descriptors, adjectives may appear as separate descriptors when designed to be precoordinated in indexing or postcoordinated in searching. They should generally not be assigned as indexing terms in isolation. Given the possibility of false coordination in searching (e.g., the linking of an adjective with the wrong noun), adjectival descriptors should be used sparingly.

Examples:

Adjectives used as descriptors	*Precoordinated indexing terms*
airborne	**airborne troops**
mobile	**mobile homes**
offshore	**offshore drilling**
portable	**portable heaters**
prestressed	**prestressed concrete**

Certain noun phrases can be used to modify other nouns, e.g., 'high frequency' can modify the noun 'waves.' The guidelines for adjectives may be applied to such noun phrases.

3.4.2.2. Economy of Cross-References

Adjectives may be used alone in general cross-references to direct the user to or from a group of descriptors beginning with a corresponding noun, e.g., "**cardiac** . . . *see also* the descriptors beginning with **heart**." An example of a reference in the opposite direction (noun to adjective) is: "**France** *see also* the descriptors beginning with **French** (French art, French language, French literature, French wines)."

This guideline applies especially when the adjective and the noun to which it is related differ widely in their derivation and spelling, for example, 'sea/marine,' 'ear/aural.' It works well in printed thesauri with a limited number of descriptors beginning with one of the terms, but explicit references from all compound forms (e.g., cardiac diseases USE **heart diseases**) may be necessary in thesauri displayed on a screen, or in printed tools in which the compound terms beginning with a particular adjective span several columns or pages, and the user may not notice the general reference.

3.4.3. Adverbs

Adverbs such as 'very' or 'highly' should not be used alone as descriptors. A phrase beginning with such an adverb may be accepted as a descriptor only when it has acquired a specialized meaning within a domain.

Examples:

very high frequency
very large scale integration
very low density lipoproteins

3.4.4. Initial Articles

The use of initial articles in descriptors should be avoided.

3.4.4.1. Deletion

Delete the initial article when the descriptor is clear without it. If not, use a parenthetical qualifier.

Examples:

arts	*rather than*	the arts
state (political entity)	*rather than*	the state

3.4.4.2. Retention

If the initial article is an integral part of a proper name, and should be searchable, it should be included in the descriptor in direct order. Otherwise, invert the article. Whether or not the initial article is included as an integral part of the name is dependent upon the language. In the following examples, English is the assumed language of the thesaurus.

Examples:
A Big Company [initialism: ABC]
Computer Place, The
El Niño
El Salvador
Los Angeles
Narrows, The

In cases where the article is included in direct order, a cross-reference from the element following the article should be provided, e.g., Salvador USE **El Salvador**.

3.5. Singular and Plural Forms

Terms can be divided into two categories: count nouns (see section 3.5.1) and noncount (mass) nouns (see section 3.5.2). The guidelines for singular and plural based on these categories apply to the formulation of both descriptors and entry terms.

3.5.1. Count Nouns

Count nouns are names of objects or concepts that are subject to the question 'How many?' but not 'How much?'. These should normally be expressed as plurals.

Examples:
books
chemical reactions
penguins
political parties
singers
vertebrates
windows

3.5.1.1. Exception to Plural Count Nouns

If in the domain of the thesaurus there is literary or user warrant for the expression of count nouns in the singular, establishment of descriptors in that form is acceptable. For example, in the field of biomedicine, the names of parts of the body are generally formulated in the singular.

Examples:
digestive system
ear
lung
nose
stomach

In a museum catalog, objects are typically treated as unique items, and descriptors are given in the singular.

Examples:
chair
oil painting
tapestry

3.5.2. Noncount (Mass) Nouns

Noncount nouns are names of materials or substances that are subject to the question 'How much?' but not 'How many?'. These should be expressed in the singular.

Examples:
copper
paint (but cf. **paints**, below)
snow
water

If the community of users served by a thesaurus regards a given substance or material as a class with more than one member, the class should be expressed in the plural.

Examples:
plastics (i.e., various types of plastic)
paints

3.5.2.1. Abstract Concepts

The names of abstract concepts, e.g., systems of belief, activities, emotions, properties, and disciplines, shall also be expressed in the singular. Some of these terms are subject to the question 'How much?'.

Examples:

beliefs	**Judaism; Taoism**
activities or processes	**digestion; distribution; migration; welding; writing**
emotions	**anger; envy; love; pity**
properties or states	**conductivity; silence; worth**
disciplines	**architecture; business; musicology; metallurgy**

3.5.2.2. Unique Entities

The names of unique entities, whether concrete or abstract, shall be expressed in the singular.

Examples:
Shangri-La
Statue of Liberty

3.5.3. Coexistence of Singular and Plural Forms

Where the singular and plural forms of a term represent different concepts, separate descriptors for each should be entered in the thesaurus as appropriate. The distinction should be indicated by a qualifier.

Examples:
bridge (game)
bridges (dentistry)
bridges (structures)

damage (injury)
damages (law)

wood (material)
woods (forested areas)

3.6. Selection of Preferred Form

The authority for the form selected should be recorded in the term record (see section 8.5). The variant forms should be included as entry terms. The source(s) of variant forms may also be recorded.

3.6.1. Usage

Descriptors should reflect as much as possible the usage of people familiar with the domain of the thesaurus. Neutral rather than pejorative or objectionable terms should be selected, e.g., **developing nations** *rather than* underdeveloped countries. See also section 3.6.4.2.

3.6.1.1. Literary Warrant

Words and phrases drawn from the literature of the field should determine the formulation of descriptors. When two or more variants have literary warrant, the most frequently used term should be selected as the descriptor.

3.6.1.2. Published Authorities

Guidance in the selection of descriptors may be found in reference works of the domain, such as dictionaries, glossaries, encyclopedias, and authoritative treatises of the field.

3.6.1.3. Other Indexing Languages

Other thesauri, abstracting and indexing services, and subject heading lists may be useful as sources for selection of preferred forms.

3.6.1.4. Personal Authorities

Opinions of subject experts regarding the preferred form of descriptors may be sought.

3.6.2. Spelling

3.6.2.1. Literary Warrant

The most widely accepted spelling of words should be adopted. If variant spellings exist and are commonly recognized, each should be entered in the thesaurus, and a cross-reference should be made from the nonpreferred to the preferred form.

Examples:

Romania	*rather than*	Roumania
theater	*rather than*	theatre

3.6.2.2. Authorities

Spelling should follow the practice of well-established dictionaries or glossaries. If a choice between spellings is made for dialectal reasons (for example, between American and British English), the choice should be adhered to consistently throughout the thesaurus; exceptions may be made for proper names.

Examples:

catalogs	*rather than*	catalogues
color	*rather than*	colour
labor	*rather than*	labour
but **Labour Party**		

3.6.3. Abbreviations, Initialisms, and Acronyms

This section deals with the choice of full vs. abbreviated forms of descriptors. Capitalization is dealt with in section 3.7.1.

3.6.3.1. Preference for Abbreviation

Abbreviations and acronyms should be selected as descriptors only when they have become so well established that the full form of the term or proper name is rarely used. Cross-references should be made from the full forms.

Examples:

AIDS (disease)	*rather than*	acquired immune deficiency syndrome
DNA	*rather than*	deoxyribonucleic acid
lasers	*rather than*	light amplification by stimulated emission of radiation
UNESCO	*rather than*	United Nations Educational, Scientific, and Cultural Organization
VTOL aircraft	*rather than*	vertical takeoff and landing aircraft

3.6.3.2. Preference for Full Form

3.6.3.2.1. General Use

The full form of terms should be selected as descriptors when the abbreviated form is not widely used and generally understood. Cross-references should be made from the abbreviated forms.

Examples:

automated teller machines	*rather than*	ATMs
driving while intoxicated	*rather than*	DWI
prisoners of war	*rather than*	POWs

3.6.3.2.2. Ambiguity

Many acronyms and abbreviations stand for more than one word or phrase; the full form of the term should therefore be selected as the descriptor in preference to the abbreviated form, even when the abbreviation has only one value in the domain of the thesaurus.

Examples:
artificial insemination *rather than* AI
artificial intelligence *rather than* AI

3.6.4. Neologisms, Slang, and Jargon

Neologisms, slang, and jargon terms often cover new concepts originating within a particular specialty, subculture, or social group. Such terms are generally not included in standard dictionaries. When no widely accepted alternative exists, the neologism, slang, or jargon term should be accepted as a descriptor.

Examples:
kludge
yuppies

3.6.4.1. Provisional Terms

A neologism, slang, or jargon term may be labeled provisional (see section 8.6.2), and may be elevated to full descriptor status as the term becomes accepted into the language.

Example:
burnout

3.6.4.2. Cross-References from Slang Terms

A slang term that is an alternative to an existing and well-established term should be admitted as an entry term to the descriptor.

Example:
psychiatrists *rather than* shrinks
[descriptor] [cross-reference]

3.6.5. Common Names and Trade Names

3.6.5.1. Preference for Common Name

A product is frequently known by a widely recognized trade name. Where an equivalent common name exists, the common name should be used in preference to the trade name, because the generic name is the more accurate representation of the concept.

Example:
diazepam *rather than* Valium

Both common and trade names may be descriptors in thesauri that cover generic products as well as specific brands, with the latter treated as narrower terms (see section 5.3.1).

3.6.5.2. Trademarks

Because trademarks are recognized by law as proprietary, they should be identified as such in a thesaurus. The identification should be an integral part of the term. One of the following designations may be used, as appropriate: (Trademark), (TM), (Reg), or (R).

Examples:
Current Contents (R)
Scotch Tape (TM)

3.6.5.3. Loss of Protected Status

A trademark may be used as a descriptor without qualification only when its legal protection has been lost.

Examples:
aspirin
escalator
thermos

3.6.6. Popular and Scientific Names

If a popular and a scientific name refer to the same concept, the form most likely to be sought by the users of the thesaurus should be chosen as the descriptor, and a cross-reference provided from the nonpreferred term. For example, **penguins** might be chosen as the descriptor in a nontechnical thesaurus, but the scientific equivalent, **Sphenisciformes**, may be selected for a zoological thesaurus.

3.6.7. Loanwords, Translations of Loanwords, and Foreign-Language Equivalents

3.6.7.1. Loanwords

Loanwords are terms borrowed from other languages that have become naturalized in the borrowing language. If such terms are well established, they should be admitted into the thesaurus. Diacritics should be included if required by the orthographic authorities for the borrowing language.

Examples:
coup d'état
gestalt
habeas corpus
pas de deux
weltanschauung

3.6.7.2. Translations of Loanwords

Occasionally a loanword and a commonly accepted translation coexist in the language of the thesaurus. The loanword should be preferred if it is more widely accepted in the domain of the thesaurus.

Example:
amicus curiae *rather than* friend of the court
[in a legal thesaurus]

If the translation becomes well established, however, it should be selected as the descriptor.

In all cases where a concept can be expressed by both a loanword and a translated equivalent, a cross-reference should be made from the nonpreferred term.

Examples:

braking radiation	*rather than*	bremsstrahlung
pen name	*rather than*	nom de plume

3.6.7.3. Foreign-Language Equivalents

Foreign-language terms, i.e., terms that have not become naturalized in the language of the thesaurus, should be linked to terms in the preferred language in cases where the foreign terms are likely to be sought by users. The language chosen for the descriptor should be that which the user would likely expect, with a cross-reference from the equivalent term in the other language. (For guidelines on multilingual thesauri, see ISO 5964.)

Example:

Shavuot	/	**Pentecost**
[Hebrew]		*[English]*

3.6.8. Proper Names

Variant forms of proper names of persons, institutions, organizations, and places, as well as titles, if uncontrolled, create problems for searchers. Frequency of need for proper name access points generally determines which of the following practices is adopted.

Proper names of persons, institutions, organizations, places, and titles:

a) may be controlled by inclusion in a thesaurus of subject (topical) descriptors;
b) may be controlled through a separate name authority file;
c) may be left uncontrolled.

In options b) or c), proper names may be assigned to documents as identifiers, differentiated from topical descriptors.

Different methods may be adopted for controlling various types of names. For example, geographic names may be included in a thesaurus of topical descriptors, and personal names established in an authority file.

Reasons for merging the files of name headings and topical descriptors are (a) that the borderline between them is not sharp, and (b) that it is often desirable to link a class with its instances hierarchically, e.g., **museums** and **Louvre** (see section 5.3.3).

When proper names are included in a thesaurus, the form of the name should be selected in accordance with a recognized code of cataloging practice, such as the *Anglo-American Cataloguing Rules*, 2nd ed.

3.6.8.1. Place Names

Names of countries and geographic regions frequently vary from language to language. Variant terms referring to the same place also occur within a single language community, for reasons such as the following:

a) An official and a popular name are both in common use:

   Example:

   Netherlands / Holland

b) Anglicized and vernacular forms coexist:

   Examples:

   Cambodia / Kampuchea

   Israel / Yisrael

The form most familiar to the users of the thesaurus should be designated as the descriptor, and cross-references should be provided from the variants. Other things being equal, preference should be given to the official rather than the popular name. The short form of the official name should be preferred. Standard authorities, such as the publications of the U.S. Board on Geographic Names, should be consulted for the official forms.

Examples:

Arabian Peninsula	*rather than*	Arabia
Netherlands	*rather than*	Holland
Philippines	*rather than*	Republic of the Philippines

3.7. Capitalization, Punctuation, and Nonalphabetic Characters

3.7.1. Capitalization

It is recommended that predominantly lowercase characters be used for descriptors in thesauri. (The display of descriptors in indexes is outside the scope of this standard.) Capitals should be used only for the initial letter(s) of proper names, trade names, and for those components of taxonomic names, such as genus, which are conventionally capitalized; for all the letters of initialisms; or where featured in unusual positions in product or corporate names. Because lowercase letters may also occur in unusual positions in proper names, using a combination of capitals and lowercase letters in thesauri indicates to the user the correct orthography of a term in natural language and serves to distinguish common nouns from similar proper names.

Examples:

dBASE IV (Trademark)
DNA
information systems
Information Systems Corp.
NewsBank

3.7.2. Nonalphabetic Characters

To simplify filing and searching, the use of symbols and punctuation marks in descriptors and entry terms should be minimized. Established orthographic authorities determine when such characters are essential.

3.7.2.1. Parentheses

Parentheses should be used only to enclose qualifiers (see section 3.2.1) and trademark indicators (see section 3.6.5.2), or when they constitute part of a term.

Example:
benzo(a)pyrene

3.7.2.2. Hyphens

To obviate problems in filing and searching, hyphens should be avoided in thesauri whenever possible. Retain hyphens in topical terms only when dropping them would lead to ambiguity.

Examples:
high temperature testing
n-body problem
nonfiction

Hyphens should be retained where they occur as part of abbreviations, trademarks, chemical names, or proper nouns.

Examples:
MS-DOS
Newton-John, Olivia
p-benzoquinone
Saur, Karl-Otto
Stratford-upon-Avon

3.7.2.3. Apostrophes

3.7.2.3.1 Possessive Case

Apostrophes indicating the possessive case, whether in the singular or plural, must be retained. (For the choice of singular or plural descriptors, see section 3.5.) The singular form is most frequently found in eponyms. Standard orthographic authorities should be consulted for the position of the apostrophe in common nouns.

Examples:

singular	*plural*
Boyle's law	artists' marks
Down's syndrome	workers' compensation
Pascal's triangle	

3.7.2.3.2. Proper Names

Apostrophes that are part of proper names must be retained.

Example: O'Dwyer's tubes

3.7.2.4. Diacritical Marks

Diacritical marks may be used if they are required for proper names or by the accepted standards of a discipline. (See section 3.6.7.1 for the use of diacritics in loanwords and section 3.7.3 for the use of diacritics in romanization.)

Examples:
El Niño
Guillain-Barré syndrome
Mössbauer effect
résumés

3.7.2.5. Other Symbols and Punctuation Marks

Symbols, such as the ampersand, and punctuation marks should be used only in trademarks and proper names because they create filing and searching problems.

Examples:
A & W Food Services
PL/C

3.7.3. Romanization

Commonly accepted spellings for terms or proper names from languages written in non-roman scripts, as found in authoritative reference sources, are preferable to systematic romanization, i.e., applying a table to convert the characters in a non-roman script into roman ones. A cross-reference should be provided from the systematic romanization to the established spelling.

Examples:

Chebyshev polynomials	*rather than* *or*	Tchebyshev polynomials Čebyšev polynomials
Koran	*rather than*	Qur'ān

When romanizing (transliterating or transcribing), the procedure described in a relevant standard or a nationally accepted library scheme should be followed as far as possible. (Library schemes used in the U.S. have been collected in *ALA-LC Romanization Tables,* 1991.) If a choice exists, a romanization system that uses few or no diacritical marks should be selected.

4. Compound Terms

4.1. General

Compound or multiword terms in natural language are considered lexemes, i.e., words bound together as lexical units. Dictionaries differ in their policies regarding the inclusion of various categories of compounds, and so they are not always authorities in this regard for thesaurus

designers. For this reason, guidelines on the handling of compound terms in thesauri have been developed.

To be acceptable as a descriptor, a compound term should express a single concept or unit of thought, capable of being arranged in a genus-species relationship within a hierarchy or tree structure. The compound phrase 'children and television,' for example, cannot be accommodated either in a strict genus-species hierarchy on 'children' or one on 'television,' while the compound term 'adopted children' represents a type of children, and the compound term 'educational television' represents a type of 'television.' (A compound phrase such as 'children and television' could be subordinated to a hierarchy on 'children and media,' but this would not represent a single genus-species relationship.)

4.1.1. Purpose of Guidelines on Compound Terms

The establishment of procedures for dealing consistently with compound terms is one of the most difficult areas in the fields of thesaurus construction and indexing. The guidelines below have the following purposes:

a) to aid in achieving intra-indexer, inter-indexer, and inter-organizational consistency;
b) to avoid overcomplexity of the indexing language;
c) to achieve a thesaurus whose structure is based upon principles that are amenable to logical explanation;
d) to enhance the ease and precision of searching.

4.1.2. Precoordination and Postcoordination

Descriptors should represent single concepts, expressed by a single word or by a multiword term. To express complex subjects, descriptors may be postcoordinated in search statements using Boolean or other operators, e.g., **liver** AND **anatomy**, or precoordinated into semantically linked, heading-subheading combinations, e.g., **liver—anatomy**.

4.1.3 Factors to Be Considered

The factors enumerated below may be considered in deciding which multiword terms should be split into separate descriptors and which should be retained in compound form. It should be emphasized that these procedures need not be applied rigidly. Administrative policies or system constraints may govern which of these practices is adopted by a given organization.

a) *Literary warrant.* A compound term may be employed so frequently within the literature of the domain covered by the thesaurus that splitting the term into its components would be unacceptable to users who consider it a lexeme.
b) *Regulating the number of terms in the thesaurus.* Compound terms increase the number of descriptors in a thesaurus.
c) *Printed versus computerized retrieval systems.* If a thesaurus is used for printed as well as electronic publications, the requirements of users of both formats should be taken into account. Printed indexes may require precoordinated headings for compound concepts, so that the number of entries per heading can readily be scanned by a searcher, whereas in computerized retrieval, access is often enhanced by the indexing of compound concepts through separate descriptors, which can subsequently be postcoordinated by the searcher. Rules may be developed for combining descriptors to form precoordinate indexing terms; while this practice reduces the size of a thesaurus, it should be kept in mind that the inclusion of compound descriptors may simplify the indexing process.
d) *Avoidance of false drops in retrieval.* Compound descriptors often obviate certain kinds of false drops (unwanted combinations of terms) in retrieval when the same words in a different sequence have a different meaning. For example, if the concept 'library science' is represented by the separate descriptors 'library' and 'science,' a search on this combination will also retrieve 'science library.'
e) *The nature of terminology in a given field.* This may call for special criteria to regulate the treatment of compound terms.

4.1.4. Elements of Compound Terms

Elements of a compound term are distinguished by their logical roles or functions, which are relevant to the criteria for determining whether a compound term should be split (see section 4.2). The majority of compound terms, including both pre- and postmodified noun phrases, can be analyzed into the following components:

a) the focus (also known as the 'head noun'): the noun component that identifies the broader class of things or events to which the term as a whole refers.

Examples:

concrete	*in the compound term*	**reinforced concrete**
glass	*in the compound term*	**stained glass**

b) the difference (also known as the 'modifier'): the part of a compound term that refers to a characteristic, or logical difference, which narrows the denotation of the focus by specifying a subclass of the broader concept represented by the focus.
Examples:
reinforced, which specifies a subclass, i.e., a type, of concrete in **reinforced concrete**
stained, which specifies a subclass of glass in **stained glass**

4.2. Criteria for Establishing Compound Terms as Descriptors

Compound or multiword terms should be established as descriptors in the following circumstances:

a) Splitting the parts would lead to ambiguity or a loss of meaning.
Examples:
plant food
rose windows

b) One component of the term is not relevant to the scope of the thesaurus or is too vague to exist as an independent term.
Examples (nondistinctive element underlined):
composite drawings
first aid
stone cutters

c) The modifier in the term has lost its original meaning, and so the meaning of the compound term as a whole is not the sum of the meaning of its parts.
Examples:
deck chairs
trade winds

d) The modifier suggests a resemblance, as in a metaphor, to an unrelated thing or event.
Examples:
butterfly valves
tree structures

e) The term contains an adjective that does not define a subclass of the focus, and is not actually a member of that class at all.
Examples:
artificial limbs
paper flowers
rubber ducks
tin soldiers

f) The term is a proper name, or includes proper nouns or proper adjectives.
Examples:
Freudian slips
Hudson's Bay Company
United Nations

g) The compound term has become so familiar in common use, or in the field covered by the thesaurus, that it is considered for practical purposes to represent a single concept.
Examples:
data processing
gross domestic product

4.3. Criteria for Determining when Compound Terms Should Be Split

4.3.1. Factors to Be Considered

It is more difficult to specify exact criteria for recognizing the kinds of compound terms that should be split into separate components, each of which is then entered as a noun or noun phrase in the thesaurus. The following recommendations are based upon general criteria, but they should not be regarded as mandatory instructions to be applied rigidly in all circumstances. As noted above, such decisions are often based on the nature of terminology in a given field, which may call for special treatment of compound terms, or on the necessity for controlling the size of the thesaurus.

4.3.2. Hierarchical Structure

The following recommendations are offered as a means of achieving consistent treatment of compound terms that should be split. They also ensure a correct hierarchical structure in which each term fits into a logical conceptual framework. Each of these recommendations is accompanied by a converse condition that can be used to identify classes of terms that can usually not be split without risk of ambiguity.

4.3.2.1. Focus and Difference

a) A compound term should be split when its focus refers to a property or part, and its difference represents the whole or possessor of that property or part.
Examples:
aircraft engines = **aircraft** + **engines**
hospital personnel = **hospitals** + **personnel**
soil acidity = **soils** + **acidity**

b) Conversely, a compound term should not be split when the focus term refers to a whole and the difference is a term for its part or property.
Examples:
acid soils
skilled personnel

4.3.2.2. Transitive Action

a) A compound term should be split if it consists of a term representing a transitive action modified

by a term for the object on which the action is performed.
Examples:
office management = **offices** [object] + **management** [action]
textile printing = **textiles** [object] + **printing** [action]

b) Conversely, a compound term should not be split when its component term for a thing or material is modified by a term for the action carried out upon it.
Examples:
cast steel (cast [action] + steel [object])
printed textiles (printed [action] + textiles [object])

4.3.2.3. Intransitive Action

a) A compound term should be split if it consists of a term for an intransitive action modified by a term for the performer (agent) of the action.
Examples:
bird migration = **birds** [agent] + **migration** [action]
metal deterioration = **metal** [agent] + **deterioration** [action]

b) Conversely, a compound term should not be split when the term for an object is modified by a term for the intransitive action in which the object is involved.
Example:
dancing shoes (dancing [action] + shoes [object])

4.4. Compound Descriptors vs. Node Labels

A compound descriptor should not be created solely for the reason that it forms a logical level in a hierarchy and would serve to group a set of narrower terms. For this purpose node labels should be used (see section 5.3.5).

4.5. Order of Words in Compound Terms

Noun phrases should be entered in a thesaurus in natural language order, i.e., direct order. For further guidance on the form of premodified and postmodified noun phrases, see section 3.4.1.2.
Examples:
birds of prey
oral surgery
secondary schools

4.5.1. Cross-References from Inversions

The inverted form of a noun phrase may be included as a cross-reference to the preferred term in natural language order.
Examples:
surgery, oral USE **oral surgery**
schools, secondary USE **secondary schools**
See also the discussion of permuted lists in section 6.1.1.

5. Relationships

5.1. General

5.1.1. Semantic Linking

One of the main differences between thesauri and other vocabulary lists is that thesauri display and clearly distinguish the basic semantic relationships that link the terms they contain by means of relationship indicators. Other types of term lists do not provide such relationship indicators, or do not provide them systematically. Ideally, a thesaurus should not include any orphan terms—descriptors that are not related to any other descriptors.

A variety of formats for displaying semantic relationships are described in section 6; this section is concerned with defining and illustrating the relationships themselves.

5.1.2. Kinds of Relationship

Relationships of three kinds should be included in thesauri:
a) the equivalence relationship (see section 5.2);
b) the hierarchical relationship (see section 5.3);
c) the associative relationship (see section 5.4).

These relationships are illustrated in Figure 1, p.14.

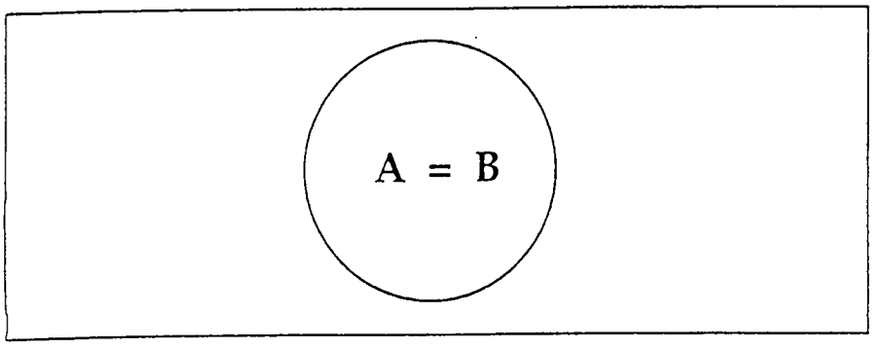
Each relationship possesses the property of reciprocity, i.e., every relationship indicated between term A and term B has a corresponding relationship from term B to term A.

The conventional abbreviations for relationship indicators are used in the examples below. Additional abbreviations for specialized purposes are found in the following sections. (A complete list of abbreviations used in this standard appears on page xii. Typographic aspects of relationship indicators are discussed in section 6.3.3.)

Relationship	Relationship Indicator	Abbreviation
Equivalence	USE	none or U
(Synonymy)	USED FOR	UF
Hierarchy	BROADER TERM	BT
	NARROWER TERM	NT
Association	RELATED TERM	RT

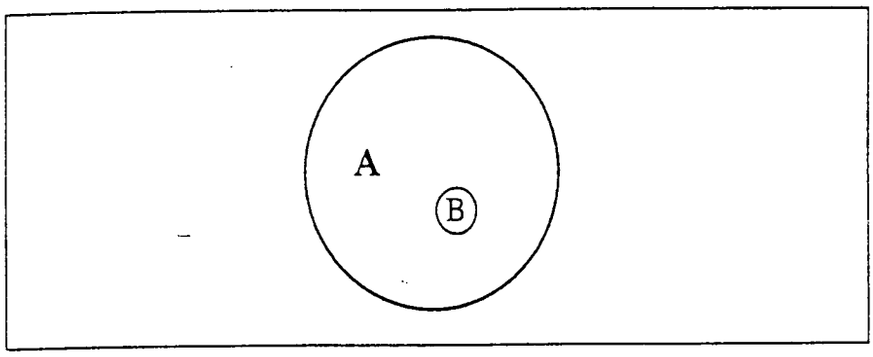
Figure 1. Semantic Relationships in a Thesaurus

The circles in the diagrams below represent the scope of the terms.



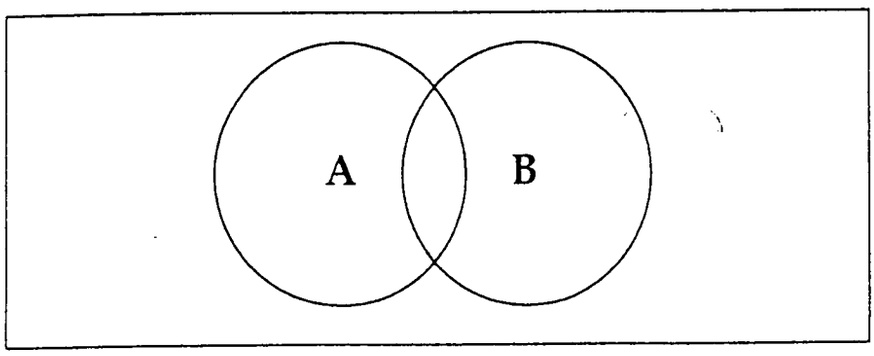
The diagram implies equivalent sets. Circle A and Circle B overlap.

Figure 1a. The equivalence relationship, e.g., teenagers USE adolescents.
(A) (B)



The diagram implies class inclusion.

Figure 1b. The hierarchical relationship, e.g., mammals NT cows.
(A) (B)



The diagram implies semantic overlap, i.e., that there is an element of meaning common to both terms.

Figure 1c. The associative relationship, e.g., gold RT money.
(A) (B)

The relationship indicators are binary (or paired) operators. Some indicators are symmetric, and some are asymmetric.

RT is symmetric:
if A RT B, then B RT A.

USE and UF are asymmetric:
if A USE B, then B UF A.

Similarly, for BT and NT:
if A BT B, then B NT A.

5.2. The Equivalence Relationship

5.2.1. General

When the same concept can be expressed by two or more terms, one of these is selected as the preferred term, i.e., the descriptor. The relationship between preferred and nonpreferred terms is an equivalence relationship in which each term is regarded as referring to the same concept. The descriptor in effect substitutes for other terms expressing equivalent or nearly equivalent concepts. A cross-reference to the descriptor should be made from any synonym or quasi-synonym that may function as an entry term for the user.

Reciprocity of the equivalence relationship is expressed by the following conventions:

U or **USE**,	which leads from a nonpreferred (entry) term to the descriptor, and
UF or **USED FOR**,	the reciprocal, which records entry terms leading to the descriptor.

Examples:

Aves **USE** **birds** outline **USE** **shape**
birds **UF** Aves **shape** **UF** outline

These relationship indicators are the equivalents of *see* and *x* (see from), respectively, in many traditional subject heading lists. This equivalence relationship covers three basic types of term:

a) synonyms (see section 5.2.2);
b) lexical variants (see section 5.2.3); and
c) quasi-synonyms (see section 5.2.4).

5.2.2. Synonyms

Synonyms are terms whose meanings are regarded as the same or nearly the same in a wide range of contexts. True synonyms are rare in natural language. Although the terms may be interchangeable in many circumstances, usage may vary as a result of such factors as level of formality, professional vs. lay context, or pejorative vs. neutral vs. complimentary connotation. The examples listed below are illustrative of the various classes of synonyms that may be encountered in practice. A slash is used to separate the synonyms, indicating that the preferred term has not been selected. (For this reason, the terms are not in boldface.)

a) terms of different linguistic origin
Examples:
cats / felines
freedom / liberty
sodium / natrium
sweat / perspiration

b) popular terms and scientific names
Examples:
aspirin / acetylsalicylic acid
gulls / Laridae
salt / sodium chloride

c) generic nouns and trade names
Examples:
petroleum jelly / Vaseline
photocopies / Xeroxes
refrigerators / Frigidaires
tissues / Kleenex
(See section 3.6.5.2 for the designation of trademarks.)

d) variant names for emergent concepts
Example:
hovercraft / air cushion vehicles

e) current or favored terms versus outdated or deprecated terms
Examples:
poliomyelitis / infantile paralysis
developing countries / underdeveloped countries

f) common nouns and slang or jargon terms
Examples:
helicopters / whirlybirds
psychiatrists / shrinks

g) dialectal variants
Examples:
elevators / lifts
subways / undergrounds

In these and other cases, preferred terms should be selected to serve the needs of the majority of users, bearing in mind the criteria enumerated in sections 3 and 4. For the sake of predictability, these criteria should be applied consistently throughout the thesaurus. If, for example, American spelling is preferred to British spelling, or scientific names are preferred to popular names, this decision should be noted in the introduction (see section 6.3.2) and should be applied consistently in the formulation of descriptors.

5.2.3. Lexical Variants

Lexical variants differ from synonyms in that synonyms are different terms for the same concept, while lexical variants are different word forms for the same expression. These forms may derive from spelling or grammatical variation or from abbreviated formats.

a) lexical variants, including direct versus inverted order, orthographic variants, stem variants, and irregular plurals

Examples:

radar antennas / antennas, radar
Romania / Rumania / Roumania
ground water / ground-water / groundwater
online / on-line
pediatrics / paediatrics
mice / mouse

b) full names and abbreviations

Examples:

International Federation for Documentation / FID
pi mesons / pions
polyvinyl chloride / PVC

5.2.4. Quasi-Synonyms

Quasi-synonyms are terms whose meanings are generally regarded as different, but which are treated as equivalents for the purposes of a thesaurus, e.g., 'REM sleep' and 'rapid eye movements.' The extent to which terms are treated as quasi-synonyms depends in large measure upon the domain covered by the thesaurus and its size.

Quasi-synonyms are frequently antonyms or represent points on a continuum:

Examples:

wetness / dryness
smoothness / roughness

An indexing agency may decide to designate either 'wetness' or 'dryness' as the preferred term on the understanding that the specialist studying one of these properties should be able to retrieve all information on the other.

As a general rule, terms should be treated as quasi-synonyms only in subject areas that are peripheral to the domain of the thesaurus. When concepts can be distinguished in the thesaurus domain with sufficient precision to justify their representation as separate terms in a thesaurus, they should be individually defined and retained as descriptors. If two concepts cannot be consistently and reliably differentiated from one another, however, a term for one concept should be selected as the descriptor and a USE reference made from the other.

5.2.4.1. Generic Posting

Generic posting is a technique in which the name of a class and the names of its members are treated as quasi-synonyms, with the broader class name functioning as the preferred term.

Examples:

waxes		plant waxes	
UF plant waxes		USE	**waxes**
furniture			
UF beds	beds	USE	**furniture**
UF chairs	chairs	USE	**furniture**
UF desks	desks	USE	**furniture**
UF tables	tables	USE	**furniture**

If employed, this technique should be limited to the peripheral areas of a subject field, or used when the number of documents on the members of a class does not warrant its being split into subclasses. This practice places limits on the specificity of the thesaurus, and should be used with discretion. The narrower terms are useful entry terms.

5.2.5. Cross-References to Elements of Compound Terms

A USE... AND (or USE+) reference may be made from a compound term to its components in cases where the term is split and both components must be used in indexing or searching.

Examples:

coal mining	USE **coal** AND (or+) **mining**
ferromagnetic films	USE **ferromagnetic materials** AND **films**

UF+ (USED FOR... AND...) is a code that may be used for the reciprocal of the reference from the compound term, as UF alone may suggest that one of the two components may be used alone in searching for the concept.

Example:

coal	UF+	coal mining
mining	UF+	coal mining

This type of cross-reference is illustrated in Figure A1.

5.3. The Hierarchical Relationship

This basic relationship is the primary feature that distinguishes a systematic thesaurus from an unstructured list of terms, such as a glossary. It is based on degrees or levels of superordination and subordination, where the superordinate descriptor represents a class or a whole, and subordinate descriptors refer to its members or parts. Reciprocity may be expressed by the following relationship indicators:

BT (Broader Term), a label for the superordinate descriptor

NT (Narrower Term), a label for the subordinate descriptor

Example:

mammals **vertebrates**
BT **vertebrates** NT **mammals**

In the flat format thesaurus (see section 6.2.1) BT and NT indicate one level broader and one level narrower, respectively. There are other types of alphabetical display that indicate multiple levels of hierarchy (see section 6.2.2).

Hierarchical relationships may also be indicated by systematic presentations such as tree structures (see section 6.2.3) or graphic displays (see section 6.2.4).

The hierarchical relationship covers three logically different and mutually exclusive situations:

a) the generic relationship (see section 5.3.1);
b) the whole-part relationship (see section 5.3.2);
c) the instance relationship (see section 5.3.3).

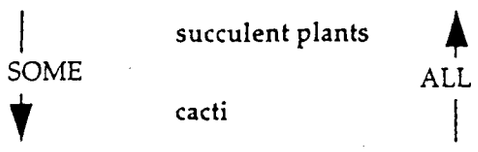
All these types of relationship may apply to a single descriptor, and special codes may be used to distinguish them. Each of the relationships leads to hierarchies that are amenable to a logical test (see below) through reference to the basic types of concept represented by the terms, for example, those listed in section 3.3. Every subordinate descriptor should refer to the same basic kind of concept as its superordinate descriptor, that is, both the broader and the narrower term should represent a thing, an action, a property, etc.

Example:

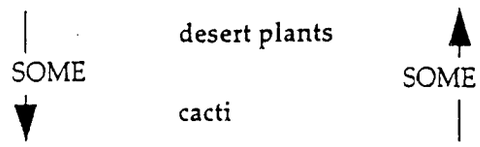
a) **anatomy** (a discipline) and **central nervous system** (a body part that may be an object of study of that discipline) represent different kinds of concept; therefore, these terms cannot be related hierarchically;
b) **central nervous system** and **brain** both represent body parts; these terms can therefore be related hierarchically.

5.3.1. The Generic Relationship

This relationship identifies the link between a class and its members or species. A simple way to apply the test for validity described above is to formulate the statement "[narrower term] *is a* [broader term]." This relationship is also amenable to a logical 'all-and-some' test as shown in the following diagram:



The diagram indicates that some members of the class 'succulent plants' are known as 'cacti' and that all 'cacti,' by definition and regardless of context, are 'succulent plants.' This test would usually ensure that a descriptor such as 'cacti' is not subordinated to a class such as 'desert plants,' since cacti are not desert plants by definition. This particular relational link can be represented by the following diagram:



The diagram illustrates that some members of the class 'desert plants' are known as 'cacti,' and that some, but not all, cacti are desert plants. These descriptors should therefore be assigned to different hierarchies in the thesaurus, and both descriptors should be assigned to the same document when indexing a work on "cacti as desert plants."

5.3.1.1. Codes for the Generic Relationship

The generic nature of a relationship may be identified by the BT/NT coding described above, or if more refined coding is desired for the various types of hierarchical relationship, by the following abbreviations:

BTG = Broader term (generic)
NTG = Narrower term (generic)

Example:

rats **rodents**
BTG **rodents** NTG **rats**

5.3.2. The Whole-Part Relationship

This relationship covers situations in which one concept is inherently included in another, regardless of context, so that the descriptors can be organized into logical hierarchies, with the whole treated as a broader term. This relationship can be applied to several types of term; the four types enumerated below are not intended to be exhaustive. In the following examples, parts are indicated through indention.

a) systems and organs of the body

Example:

nervous system
  central nervous system
    brain
    spinal cord

b) geographic locations

Example:

Canada
  Ontario
    Ottawa
    Toronto

c) disciplines or fields of discourse
Example:
science
biology
botany
zoology

d) hierarchical organizational, corporate, social, or political structures
Examples:
countries
states/provinces
cities
armies
divisions (military)
battalions
regiments

5.3.2.1. Codes for the Whole-Part Relationship

The hierarchical whole-part relationship may be indicated specifically by the following abbreviations:

BTP = Broader term (partitive)
NTP = Narrower term (partitive)

Example:

central nervous system	**nervous system**
BTP **nervous system**	NTP **central nervous system**

5.3.2.2. Parts of Multiple Wholes

When a whole-part relationship is not exclusive to a pair of terms, i.e., the part can belong to multiple wholes, the name of the whole and its part(s) should be linked associatively rather than hierarchically in the thesaurus (see section 5.4). Carburetors, for example, are parts of machines other than cars. Therefore, 'cars' and 'carburetors' should not be linked in a whole-part relationship in a thesaurus. See section 4.3.2.1 for further guidance on this point.

5.3.3. The Instance Relationship

This relationship identifies the link between a general category of things or events, expressed by a common noun, and an individual instance of that category, often a proper name.

Example:

mountain regions	*-class-*	**state capitals**
Alps	*-instances-*	**Albany**
Himalayas		**Trenton**

In the first example, the **Alps** and the **Himalayas** are assigned to subordinate positions in a hierarchy, yet they are neither kinds nor parts of **mountain regions**, but represent specific examples or instances.

5.3.3.1. Codes for the Instance Relationship

The hierarchical instance relationship may be indicated specifically by the following abbreviations:

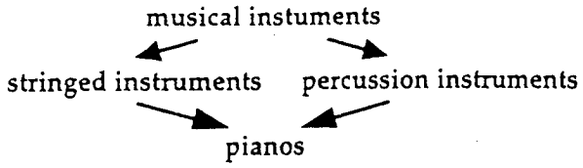
BTI = Broader term (instance)
NTI = Narrower term (instance)

Example:
fairy tales
NTI **Cinderella**
Rumpelstiltskin

5.3.4. Polyhierarchical Relationships

Some concepts belong, on logical grounds, to more than one category. They are then said to possess polyhierarchical relationships.

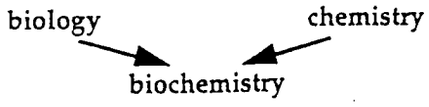
Example:



In this example, the term **pianos** is assigned to subordinate positions on the basis of its generic relationship to two broader terms; in other words, **pianos** would be an NT to both **stringed instruments** and **percussion instruments**.

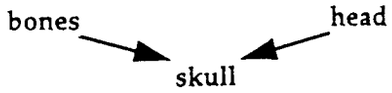
In other cases, polyhierarchical links may be based upon whole-part relationships.

Example:



In some cases, polyhierarchical links can be based on logically different relationships.

Example:



In this example, the link between **bones** and **skull** is based upon the generic relationship (the skull is a kind of bone), whereas the link between **head** and **skull** is based on the hierarchical whole-part relationship (the skull is part of the head). These relationships may be represented as

skull
BTG **bones**
BTP **head.**

5.3.5. Node Labels in Hierarchies

When descriptors are arranged in hierarchies (such as tree structures) in a thesaurus (see section

6.2), node labels may be inserted into arrays to show the principles of division among a set of sibling terms (descriptors that share a broader term). Although their function is similar to that of BTs, *node labels are not descriptors*, and must not be used as indexing terms. They are often typographically distinguished from descriptors, e.g., through the use of italics and/or enclosure in angle brackets.

Example:

cars
by motive power
diesel cars
electric cars
by purpose
racing cars
sports cars

Additional examples of node labels are in Figure A8 and Appendix B. For the use of node labels in the associative relationship, see section 5.4.3.

5.4. The Associative Relationship

This relationship covers associations between descriptors that are neither equivalent nor hierarchical; yet the terms are semantically or conceptually associated to such an extent that the link between them should be made explicit in the thesaurus, on the grounds that it may suggest additional descriptors for use in indexing or retrieval. The relationship is symmetrical, and is generally indicated by the abbreviation RT (related term). (Cf. section 5.4.4.)

Example:

cells	cytology
RT cytology	RT cells

The associative relationship is the most difficult one to define, yet it is important to make explicit the nature of the relationship between descriptors linked in this way and to avoid subjective judgments as much as possible; otherwise, RT references may be established inconsistently.

As a general guideline, whenever one term is used, the other should always be implied within the common frames of reference shared by the users of the thesaurus. Moreover, one of the terms is often a necessary component in any explanation or definition of the other; the term 'cells,' for example, forms a necessary part of the definition of 'cytology.'

Either of the following types of terms can be linked by the associative relationship:

a) those belonging to the same hierarchy (see section 5.4.1);
b) those belonging to different hierarchies (see section 5.4.2).

5.4.1. Relationships Between Descriptors Belonging to the Same Hierarchy

5.4.1.1. Relationships Between Overlapping Sibling Terms

RT references are required for sibling terms with overlapping meanings, such as 'ships' and 'boats,' where each of the terms can be precisely defined (so they do not form an equivalence set), yet they are sometimes used loosely and almost interchangeably; the user interested in one should therefore be reminded of the other.

This link does not need to be made explicit in the systematic section of a thesaurus containing organized hierarchies (see section 6.2.3), since this method of display automatically brings such terms together. The link should be indicated, however, in an alphabetical thesaurus, and in the alphabetic section of a hierarchical thesaurus.

Example:

boats	ships
BT vehicles	BT vehicles
RT ships	RT boats

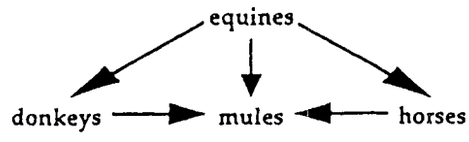
5.4.1.2. Relationships Between Mutually Exclusive Sibling Terms

It is not necessary to interrelate all sibling terms. For example, there is no need to associate terms such as 'roses' and 'daffodils,' which share the broader term 'flowers,' because the meaning of the terms does not overlap, i.e., they are mutually exclusive.

5.4.1.3. Derivational Relationships

Concepts linked by a familial or derivational relationship (i.e., one of the concepts is derived from the other) also require RT references. This guideline applies to the relationship between such terms as 'parents' and 'children' or to the relationship of 'mules' to 'donkeys' and 'horses.' In the graphic display that follows, **donkeys**, **horses**, and **mules** are all subclasses of **equines**. In the alphabetic display, RT references are provided between **mules** and **donkeys** as well as **mules** and **horses**. **Horses** and **donkeys** are not linked by RTs because they do not share a derivational relationship.

Example (graphic display):



Example (alphabetic display):

donkeys	horses
BT equines	BT equines
RT mules	RT mules
equines	mules
NT donkeys	BT equines
NT horses	RT donkeys
NT mules	RT horses

5.4.2. Relationships Between Descriptors Belonging to Different Hierarchies

It is possible to establish many grounds for associating descriptors belonging to different hierarchies. Related term references are often made between etymologically related descriptors, i.e., that contain the same root, but which do not represent the same kind of thing. The following are some representative examples of typical relational situations. For guidance on coding the precise nature of a relationship, see section 5.4.4.

a) Disciplines or fields of study and the objects or phenomena studied, or the discipline's practitioners

Examples:

mathematics	mathematicians
RT mathematicians	RT mathematics
neurology	nervous system
RT nervous system	RT neurology
botany	plants
RT plants	RT botany

b) Operations or processes and their agents or instruments

Examples:

temperature control	thermostats
RT thermostats	RT temperature control
hunters	hunting
RT hunting	RT hunters

c) Objects or processes and their counteragents

Examples:

plants	herbicides
RT herbicides	RT plants
inflammation	anti-inflammatory agents
RT anti-inflammatory agents	RT inflammation

d) Actions and their products

Examples:

weaving	cloth
RT cloth	RT weaving
lacrimation	tears
RT tears	RT lacrimation

e) Actions and their targets

Examples:

harvesting	crops
RT crops	RT harvesting
binding	books
RT books	RT binding

f) Objects or substances and their unique properties

Examples:

poisons	toxicity
RT toxicity	RT poisons
liquids	surface tension
RT surface tension	RT liquids

g) Concepts linked by causal dependence

Examples:

death	bereavement
RT bereavement	RT death
infections	pathogens
RT pathogens	RT infections

h) Concepts and their units or mechanisms of measurement

Examples:

electric current	amperes
RT amperes	RT electric current
temperature	thermometers
RT thermometers	RT temperature

i) Phrases in which the noun is not a true broader term (see section 4.2e). For example, a 'rubber duck' is not a 'duck.'

Examples:

ducks	rubber ducks
RT rubber ducks	RT ducks
fishes	fossil fishes
RT fossil fishes	RT fishes

5.4.3. Node Labels for Related Terms

In order to bring closely related concepts together in the alphabetical array under a given descriptor, related terms may be divided into categories that do not form part of a logical hierarchy. These related terms should then be identified by a node label.

Example:

books
RT
<operations>
binding
printing

In this example, *operations* functions as a node label.

For the use of node labels to indicate principles of division for logical hierarchies, see section 5.3.5.

5.4.4. Specifying Types of Related Term References

In certain thesauri, it may be considered desir-

able to refine related term references in order to make the nature of the relationships explicit. Codes for such relationship indicators and their reciprocals may be developed locally. (See, for example, Figure A22.) These local codes should be clearly explained and illustrated in the introduction (see section 6.3.2) or documentation (see section 7.4) to the published or machine-readable thesaurus.

6. Print Display

The three basic types of format in a printed thesaurus are: (a) alphabetical, showing all the immediate relationships of each term; (b) permuted or rotated, giving access to every word in each descriptor and entry term; and (c) hierarchical, showing a display of all levels of hierarchies. Various versions of each format are illustrated by sample pages from thesauri in Appendix A.

A major principle in the design of the format of a printed thesaurus is the minimization of double lookup, i.e., the need to consult more than one sequence.

The most commonly used methods of presenting equivalence relationships and term hierarchies are described and evaluated in sections 6.1 and 6.2, respectively. A discussion of general format considerations follows in section 6.3.

6.1. The Equivalence Relationship

USE references from nonpreferred terms should be incorporated into the main alphabetical listing of a thesaurus rather than being relegated to an auxiliary "access vocabulary" or separate list of entry terms.

6.1.1. Inverted USE References and Permuted Lists

Some thesauri include an auxiliary permuted or rotated list in KWIC or KWOC format that gives access to every word in descriptors and USE references (see Figures A2 and A3). The permuted list should not be used as a substitute for the inclusion of useful inversions as USE references in the main alphabetic list.

For example, it is conceivable that in a thesaurus in the domain of library and information science, a user will search the term 'automation, library.' If the preferred term is **library automation**, a USE reference is essential. A reference from the inverted form of the term **information science** is, in contrast, not essential and would, moreover, be misleading as it might be interpreted as 'science information.' An automatically generated permuted display will, however, display the term **information science** under **science** as well as under **information**.

6.1.2. Inverted USE References and Hierarchical Display

The choice of format for the display of hierarchical relationships (see section 6.2) affects decisions on the necessity for USE references. For example, if narrower terms are included in the alphabetic display in a case such as

libraries
NT **academic libraries**
public libraries
special libraries

the following USE references may be unnecessary:

libraries, academic
libraries, public
libraries, special.

If, however, **academic libraries** has the NTs **college libraries** and **university libraries**, in the flat format (see section 6.2.1), these would not be visible under **libraries**; therefore, USE references from inversions of all the compound terms should be made. Moreover, where a descriptor has numerous BTs, NTs, and RTs, an inverted USE reference may be easier to find than the desired narrow term.

If hierarchical information is given only by a tree structure (see section 6.2.3), USE references from inversions would be desirable within the main alphabetic sequence of the thesaurus.

6.1.3. Juxtaposition of Terms

Juxtaposition of terms also plays a role in the creation of USE references for a printed thesaurus. For example, an entry term that would immediately precede or follow the descriptor to which it leads may be suppressed. (This guideline would not apply to thesauri for databases searched by computer — see section 7.3.1.) Some singular and plural forms of terms may be widely separated in a printed alphabetic list, however, and a USE reference from one to the other would then be warranted, even if the policy regarding preference for singular or plural is stated in the introduction to the thesaurus (see section 6.3.2). In the following example, several terms occur between the singular and plural forms of 'cat,' and so a reference from one to the other is desirable, especially because the singular form is the first word of a compound descriptor.

Example:
cat USE **cats**
cat fleas
catalogs
catharsis
cats

The addition of descriptors that file between the position of a potential entry term and its corresponding descriptor, or the interfiling of descriptors with uncontrolled terms in an index may create a need for additional USE references.

Similar considerations apply to the establishment of related term references in a printed thesaurus.

Most printed thesauri are produced from machine-readable files that also generate a screen display. Because a screen displays less information than a printed page, a more generous entry vocabulary is required on a screen than on a page. The reader's eye takes in the larger panorama of the printed page, and it is easier to locate descriptors that are variants of the search term. A possible approach is to include all potentially useful references in the machine-readable file and suppress unnecessary ones in the printed product (see section 7.3.1).

6.2. Hierarchical Relationships

The hierarchy of descriptors may be displayed in a variety of ways. This section explains the advantages and disadvantages of the most commonly used formats for displaying hierarchical relationships among descriptors in print. Indention should be used as a visual cue for hierarchical levels in all the formats.

6.2.1. Conventional Flat Thesaurus Structure

The most commonly used thesaurus format arranges all descriptors in alphabetical order and gives for each some or all of the following: scope notes, used for (UF) references, terms that are one level broader (BT), terms that are one level narrower (NT), and related terms (RT). This is the recommended order of relationships. Narrower terms are generally arranged alphabetically, but node labels may be used to group both narrower and related terms in categories (see section 5.3.5 and Appendix B).

Indention should be used as an additional visual cue for narrower terms.

Example:

vertebrates
NT **mammals**
reptiles

An example of the flat format is found in the ERIC thesaurus (see Figure A4).

6.2.2. Multilevel Hierarchy Within Alphabetic Display

This type of display format saves the time of the thesaurus user in that the broadest and most specific descriptors relating to the sought term are immediately identified, whereas in the flat format, the user must proceed one step at a time, going from descriptor to descriptor, to examine the full hierarchy. In both formats described in the following sections, the sibling terms of the sought descriptor can be identified only by going to the broader term. Because the hierarchy is repeated over and over again in these formats, they are far more space-consuming than the flat format and hence less desirable for a printed thesaurus. If the hierarchies are not very deep, e.g., only three or four levels, the extra space occupied may not be excessive.

6.2.2.1. Multilevel Broader and Narrower Terms

This format differs from the flat format in that it gives *all* levels of broader and narrower terms within the alphabetic display. It employs special notation such as BT1, BT2 (one level broader, two levels broader) and NT1, NT2 (one level narrower, two levels narrower). A model for this format is the *International Energy Subject Thesaurus* (see Figure A5).

6.2.2.2. Generic Structure

Multiple levels of hierarchy may be indicated without BT/NT notation, by using the abbreviation GS (generic structure) with indention and punctuation marks such as period and colon as cues to the levels of hierarchy. A model for the latter system is found in the *NASA Thesaurus* (see Figure A6).

6.2.3. Tree Structure

In this format, each descriptor is assigned a classification notation or line number, and this leads the user from the alphabetic display to the full hierarchical display (sometimes called 'systematic display' or 'classified display') in a separate sequence. *Medical Subject Headings* (MeSH) serves as a model for the format featuring classification notation (see Figure A7). The *Art and Architecture Thesaurus* provides a model of a hierarchical display with line numbers (see Figure A8), although the system also employs classification notation.

There is no redundancy in this method of thesaurus display, and it is economical of space. The panorama of the complete hierarchy is a distinct advantage to the user.

Another advantage of the tree structure is that the complete array of classes and subclasses serves as a check on the logical consistency of the

hierarchies. In the flat format or generic structure, hierarchical relationships may be built from the bottom up; in a tree structure, they are often built from the top down.

The power of a tree structure may be fully exploited in online searching (see section 7.3.3). It is, however, possible for a computer to build a complete hierarchy from a thesaurus employing a flat format or generic structure.

6.2.3.1. Notation

If a hierarchical classification scheme is applied to a tree structure, its notation must be carefully developed so that it will be hospitable to interpolation at any level. Computer-generated or humanly-assigned line numbers may be easily revised when descriptors are added, but the notation is not expressive, i.e., it does not reflect the levels of hierarchy.

Pure notation, consisting of either letters or numbers, has a smaller base than mixed or alphanumeric notation and is therefore less hospitable to new descriptors. Mixed notation may, however, create problems for novice users, who may have difficulty locating the codes. This suggests an advantage of the generic structure: all levels of broader and narrower terms are displayed to the user within the alphabetic list, and no second lookup is required. Sibling terms are not displayed in the generic structure, however, while they are an important feature of the tree structure.

6.2.4. Graphic Displays

Graphic displays are logically equivalent to a tree structure, but may not have a notation (see Figure A9). They are difficult to produce and maintain, as well as being wasteful of space. A simple graphic display may make it easier to take in the relationships of a descriptor at a glance. If too many related terms and levels of hierarchy are included, however, the display becomes difficult to comprehend.

6.2.5. The Top Term Structure

In this format, the alphabetic display includes all the relationships found in the flat format, with the addition of a relationship indicator for the top term (TT) of the hierarchy. The latter is equivalent to the broadest term in a generic structure. The top term leads the user to a separate sequence of the thesaurus in which top terms are arranged alphabetically, with each top term followed by all its narrower terms, arranged hierarchically on various levels. Since it does not include a classification notation, a top term structure does not provide guidance to the location of a descriptor within the array under its top term. The *Inspec Thesaurus* provides a model for this format (see Figures A10 and A11).

This format is more economical in terms of space than the generic structure, but like the tree structure, requires a second lookup.

6.2.6. Two-Way Hierarchical Structure

This is generally appended to a flat format thesaurus. Each descriptor is an access point, and all levels of broader and narrower terms are displayed, generally without notation and with indention as a cue to hierarchy. A model of this is found in the ERIC thesaurus (see Figure A12). This format is logically equivalent to the generic structure, i.e., in a single display one can view all of the broader and narrower terms of a descriptor. In combination with the flat format, the two-way hierarchical structure requires more space than the top term structure.

6.2.7. Broad Categories

The alphabetic arrays of some thesauri include numbers that identify the broad category to which each descriptor belongs. The notation for the category may be searchable online or used for purposes of selective dissemination of information. Such thesauri generally feature a separate section that displays the descriptors for each numbered category or subcategory in an alphabetic sequence, undifferentiated in terms of hierarchical levels. An example of such category numbers may be found in the TEST thesaurus (see Figure A13).

6.3. General Format Considerations

6.3.1. Components of a Printed Thesaurus

Given the variety of displays available, it is not possible to stipulate a standard form of layout for a printed thesaurus. Minimally, every thesaurus should include an alphabetic display.

Every published thesaurus should include a title page, a table of contents identifying all the sections, and an introduction (see section 6.3.2).

6.3.2. Introduction

All printed thesauri should contain a comprehensive introduction that clearly states

a) the purpose of the thesaurus;
b) the scope, i.e., the subject field(s) covered, with core and fringe areas separately identified;
c) the meaning of all conventions, abbreviations,

and any punctuation marks used in nonstandard ways;

d) the rules and authorities adopted in selecting the preferred forms of descriptors, and in establishing their relationships;

e) whether the thesaurus complies with a national or international standard for thesaurus construction;

f) the filing rules employed, citing an appropriate national standard when used;

g) the total number of terms, with separate totals for descriptors and entry terms (the number of descriptors indicates the size of the indexing/searching vocabulary; the ratio of entry terms to descriptors indicates the accessibility of the thesaurus to end-users);

h) the date on which the computer file for the thesaurus was put into final form before printing, which may be the date of the last update;

i) a statement on the updating policy, and the name and address of the responsible organization to which comments and suggestions should be sent.

6.3.3. Typography

Descriptors, nonpreferred terms, relationship indicators, and textual notes should be typographically distinguished. Suggested typographic specifications are: lightface or *italics* for nonpreferred terms, all capitals for relationship indicators such as USE, and **boldface** for descriptors.

Example:

teenagers USE **adolescents**

6.3.4. Capitals / Lowercase Letters

The thesaurus should serve as an orthographic authority in addition to noting preferred terminology. A combination of capitals and lowercase letters should therefore be used in thesaurus terms. Topical descriptors and entry terms should be in lowercase; proper names should be capitalized as in standard usage. Initialisms are generally all uppercase; acronyms may have an initial capital followed by lowercase letters.

6.3.5. Filing

This section provides general rules for the filing of alphabetic characters and numerals in thesauri. Guidance on specific filing problems may be found in *ALA Filing Rules* (1980), *Library of Congress Filing Rules* (1980), and *British Standard Alphabetical Arrangement and the Filing Order of Numerals and Symbols* (BS 1749:1985). These filing codes are, however, not compatible with each other in certain details, and only one of them should be chosen as an authority.

6.3.5.1. Alphabetic Characters

Terms consisting of letters should preferably be filed word-by-word rather than letter-by-letter. In word-by-word filing, a space is significant; this filing principle is called "nothing before something." It juxtaposes terms that begin with the same word.

Examples:

Word-by-word filing	*Letter-by-letter filing*
gas USE **gases**	gas USE **gases**
gas coolants	**gas coolants**
gas welding	**gases**
gases	**gaskets**
gaskets	**gas welding**

Another advantage of word-by-word filing is that the space is a significant character in computer programs that may be used to sort terms and to print the thesaurus.

A disadvantage of word-by-word filing is that it separates agglutinated compounds from those consisting of two words, e.g., 'folk song, folk tune, folklore, folktale.' USE references from the variant spellings may compensate for this problem.

6.3.5.2. Numerals

Some operating systems file numerals at the beginning of the alphabet and some at the end. Either position is acceptable, provided that the filing position of numerals is explained in the introduction to the thesaurus.

Applying the principles of ascending order and word-by-word filing to numerals, a computer will file 19 before 2, and 2001 before 30. A special sort routine is required to get the computer to file on arithmetical value, e.g., 1, 1.5, 10, 151, 1000. Special sorting instructions must be inserted by humans if numbers are to be ignored, e.g., in chemical terms, or filed as spelled out, e.g., **21** [twenty-one] **gun salute.**

6.3.5.3. Non-Alphanumeric Characters

Punctuation marks should generally be ignored in filing entry terms and descriptors. In the following example, the comma in the inverted USE reference is ignored:

cat breeds
cat, Siamese
USE **Siamese cats**
catalogs

Parentheses within terms should be ignored in filing, e.g., *benzo(a)pyrene* should be treated as if it

were spelled 'benzoapyrene.'

Parentheses around qualifiers require special treatment. It is desirable for the user to see homographic terms juxtaposed to assist in the selection of the desired term. Ignoring the parentheses around qualifiers would lead to the dispersal of homographic terms among compound terms. The desired sequence is illustrated in the following example:

Earth (planet)
earth (soil) USE **soil**
earth science

6.3.6. Running Heads

Whenever a printed thesaurus includes multiple sequences, e.g., an alphabetic list, a rotated list, and a hierarchical display, each page should feature a running head to identify the sequence. Running heads are useful even if the sequences are in separate volumes.

6.4. Multiple Versions of Printed Thesauri

Thesauri may be produced in various editions: a basic list of terms, references, and relationships designed for the end-user or occasional searcher, and a more complete version designed for the indexer and the expert searcher, which may include detailed scope notes, indexing instructions, information on term history, and postings data.

A model for multiple print formats may be found in the two editions of *Medical Subject Headings:* MeSH as found in *Index Medicus* (also known as "Black and White" MeSH), which is designed for end-users, and the *Annotated Alphabetic List,* which is designed for indexers and searchers (see Figures A14 and A15).

7. Screen Display

7.1. User Categories

The design of screen displays should take into account the needs of each anticipated class of user:

a) *thesaurus maintainers* who have expertise in indexing and thesaurus construction, and are likely to be experts in the subject domain of the thesaurus. They must have access to all views of a thesaurus and complete information about each term, with the ability to edit and manipulate term records, cross-references, classification notation, and hierarchies. They require "housekeeping" displays not needed by end-users of a thesaurus. (See section 10 for specifications of thesaurus management software.)

b) *expert searchers and indexers* who have expertise in indexing, online information retrieval, use of thesauri, or all of these. Indexers are likely to have expertise in the subject domain of the thesaurus; expert searchers may or may not have such expertise. These sophisticated users require the ability to search and view cross-references, definitions, and notes for descriptors as well as various levels of the classification or hierarchies. Postings data are especially important for searchers (see section 7.3.1). Sophisticated thesaurus displays and terminology are appropriate for these users.

c) *end-users* who are not likely to be experienced in the jargon and complexities of online information retrieval, but who can be expected to have expertise in the subject field, and to understand its terminology. Abbreviations should be avoided, or explained in a window or at the bottom of the screen on which they appear. The types of displays available to expert searchers can be useful to end-users as well, when designed with their needs in mind. End-users may benefit from on-screen instructions in addition to any printed documentation that may exist. It is important to take the needs of each anticipated class of user into account in designing screen displays.

7.2. Factors to Be Considered

A thesaurus is generally used in the context of an information storage and retrieval system, and the choice of screen display formats will depend upon the domain, the searching behavior of users in that domain, and the overall design of the information system of which the thesaurus is a component.

Viewing information on a screen differs from viewing printed information: with a screen, it is harder to browse and remember one's context; the screen is more difficult and tiring to view than printed media; and the available screen "page" size can make it difficult to grasp information that is perfectly comprehensible in printed form.

A great deal of research has been done and is ongoing concerning user search behavior and cognitive learning styles in the context of interactive search systems, library catalogs, videotex, and hypermedia. In addition to existing standards in human-computer interaction, working groups are producing draft standards for user interfaces, screen layouts, menus, windows, and other interface formats. These standards should be applied rigorously to all user interface designs. See Appendix C for sources of

standards and references to authoritative texts on interface design.

7.3. Types of Display

The following types of display are recommended for consideration (see also section 6). The selection will depend on the requirements of a given system and its users.

7.3.1. Alphabetical Listing

When a user enters a search term—whether it is a descriptor, entry term, or term not in the thesaurus—the system should display an alphabetical sequence of terms preceding and following the search term. This list should contain both descriptors and entry terms with USE references to descriptors.

Spelling variations, common misspellings, and common typographical errors should be stored as USE references in the machine-readable thesaurus to enhance retrieval, but need not be displayed to users (see section 10.2).

Thesauri linked to databases may also display the number of postings for assigned descriptors. The postings data in a thesaurus display should not include free-text occurrences of single-word descriptors.

7.3.2. Permuted Displays

Rotated or permuted listings (e.g., KWIC and KWOC displays) of multiword terms are useful to all classes of users (see Figures A16 and A17).

7.3.3. Hierarchies and Classifications

It should be possible to display a hierarchy at various levels. For example, if the hierarchy of a full thesaurus has five levels, it should be possible to display an outline of the first three levels. Indention or other cues to the hierarchy are essential.

Users should be able to broaden or narrow their searches by selecting descriptors at various levels in the hierarchy, and to add some or all broader or narrower terms for a descriptor. For example, beginning with the descriptor **vertebrates**, the user may request that the search be expanded to include all of its narrower terms—at one or more levels—or only **mammals** and, optionally, its narrower terms. A tree structure with notation facilitates such search expansion, although the same effect may be achieved with other structures. Alternatively, the system could assign to each descriptor in a hierarchical display a line number, to which the user may refer (rather than rekeying terms) in expanding a search. Highlighting is another option to obviate rekeying.

Node labels in hierarchies, which are not to be used in indexing or searching, should be distinguished from descriptors, e.g., by placing them in angle brackets. Node labels may also be assigned line numbers.

7.3.4. Term Detail

A user should be able to choose a term from any listing and see an expanded view of the detail for that term, either fully or in part. Users should have the option of calling up a descriptor's history, scope note, or definition, as well as all term relationships —equivalence, hierarchical, and associative—plus any specialized relationships created for the thesaurus. These relationships should be appropriately labeled, avoiding abbreviations in the case of end-user displays (see Figure A18). Windowing techniques may be employed effectively in term-detail screens.

7.3.5. Graphic Displays

Research has shown that graphic displays may communicate relationships among concepts more effectively than linear displays to some users. Graphic displays of thesauri should be considered, taking into account the domain and search habits of users. (Cf. section 6.2.4.)

7.4. Documentation

For publicly available databases, thorough descriptions of the thesaurus and its use should be available in the database documentation. The basic information sheets provided by database vendors should explain the structure of the thesaurus and the commands for searching it. Thesaurus software packages should provide clearly written, well-organized, and indexed documentation for thesaurus maintainers, as well as for indexers and searchers where the thesaurus is part of a complete information retrieval package.

7.5. Relation of Thesaurus Character Set to Search Commands

Nonalphabetic characters used in a machine-readable thesaurus should not conflict with special characters used in search commands. For example, parentheses used in qualifiers should not be interpreted as nesting indicators in a search statement. If the potential for ambiguity exists, a different character should be substituted, e.g., square brackets may enclose qualifiers.

8. Thesaurus Construction

8.1. Avoidance of Duplicate Work

The compiler should ascertain, through reference to one of the clearinghouses listed in Appendix D, whether an existing thesaurus covers the same or an overlapping domain of knowledge. Complete duplication of subject coverage is rare, but access to one or more thesauri in related fields can frequently serve as a useful starting point. For example, a thesaurus developer in the field of malpractice law should begin by consulting existing thesauri in medicine and law, which will surely contain many of the terms and relationships required for the more specialized, interdisciplinary thesaurus.

8.2. Determination of Structure and Display Format

If possible, the form of the thesaurus (e.g., flat, generic structure, hierarchical display, or graphic display) should be decided before terms are collected and considered as candidates for inclusion. As demonstrated in section 6, the display format of the thesaurus affects the types of cross-references and relationship indicators that are provided. For example, if a tree structure with classification notation is adopted, the thesaurus may not include BT/NT references. The format should therefore be selected before relationships among terms are constructed. (Although it is possible to convert one thesaurus display format to another, one must begin with consistent coding of relationships in a single system.)

8.3. Methods

Three initial approaches to thesaurus construction are recommended.

8.3.1. The Committee Approach

Experts in the subject domain of the thesaurus draw up a list of the key terms in the field and indicate the relationships among them, with assistance from experts in thesaurus design.

8.3.2. The Empirical Approach

a) *The deductive method.* Terms are extracted from documents (by humans or computers; see section 8.4 for a discussion of machine assistance), optionally during a preliminary stage of indexing, but no attempt is made to control the vocabulary, nor to determine relationships between terms, until a sufficient number of terms has been collected. All terms are then reviewed by a group of experts, preferably consisting of both information and subject specialists. They should first identify terms that represent the broadest classes, and then allocate remaining terms to these classes on the basis of their logical relationships, so that the hierarchies tend to be established on a broader-to-narrower basis. Vocabulary control should be applied at the stage where hierarchies are established, following the principles described in sections 3 and 4.

b) *The inductive method.* New terms are selected for potential inclusion in the thesaurus as they are encountered in documents. Vocabulary control is applied from the outset. Each term, as it is admitted, is designated as a member of one or more broader classes that are constructed on an ad hoc basis at an early stage. The thesaurus is therefore established on a narrower-to-broader term basis. Thesaurus construction is regarded from the outset as a continuous operation. Assistance from subject experts is strongly recommended; these experts may serve as members of a formal editorial board or committee.

8.3.3. Combination of Methods

In practice, more than one of these approaches is likely to be employed at one stage or another during the construction of a thesaurus. For example, hierarchies of terms that were first established inductively may later be examined from a deductive viewpoint. Both techniques are essentially empirical, and it should be accepted from the outset that some decisions regarding the terms and their interrelationships that were made during the early stages of compilation may have to be revised as further experience is gained.

The compilers should check terms and hierarchies frequently to ensure consistent application of principles in such procedures as the construction of inter-term relationships and the splitting of compound terms.

8.4. Machine Assistance

It is assumed throughout this standard that thesaurus construction calls for intellectual decisions. Machine assistance can be employed, however, for term identification tasks such as the following:

a) *Identification of candidate terms.* Candidate terms may be identified automatically from machine-readable text, e.g., titles and/or abstracts. The number of potential terms should be reduced by use of a stop list (a list of function words—articles, conjunctions, and prepositions such as 'a,' 'and,' and 'for'—plus other words consid-

ered to be of no value for retrieval). Remaining terms should be matched against those already recorded in the thesaurus. Unrecorded terms may be considered as candidates for inclusion.

b) *Registering frequency of descriptor assignment.* In computerized indexing systems, the frequency with which a descriptor has been used in indexing can be registered automatically. Descriptors with exceptionally high or low scores can be considered as candidates for modification or deletion.

c) *Recording terms in user queries.* Terms found in user queries that do not match one or more descriptors or entry terms may also be considered for inclusion, especially when a given term occurs in multiple queries. To ensure privacy, users should not be identified.

8.5. Term Records

An individual record should be created for every descriptor, and optionally for every entry term, as soon as it is admitted into a thesaurus. Records for entry terms may include source notes as well as the date of admission into the thesaurus. For descriptors, the record may contain any or all of the following elements:

- descriptor;
- source(s) consulted for descriptors and entry terms. This field is especially important for neologisms or unfamiliar terms; it may include citations to published sources or the names of personal authorities consulted;
- scope note;
- synonyms, i.e., Used For references;
- nondisplayable variations, e.g., common spelling errors (see section 10.2);
- broader terms;
- narrower terms;
- related terms;
- locally established relationships (see, for example, Figure A22);
- category or classification number;
- history note (generally coded HN), including minimally the date added, as well as the record of changes, if any (see section 9.3). If appropriate, the history note may include the date discontinued, the term that succeeded the descriptor, and/or the term that preceded it.

An example of a term record is in Figure A18.

Section 10.10 discusses field definition in thesaurus management systems.

8.6. Term Verification

Before a descriptor is admitted into a thesaurus, it should be validated according to the rules in section 3 concerning scope, form, and choice of descriptors. The compiler should also review relationships between each new descriptor and other descriptors in the hierarchy to which it is assigned. The following types of authority within the subject domain should be checked before candidate terms are accepted for inclusion:

a) technical dictionaries, glossaries, scholarly monographs, reference texts, and encyclopedias;
b) existing thesauri;
c) classification schemes.

In problematic cases, expert advice should be sought on the selection of a descriptor from variant forms of terms.

8.6.1. Candidate Terms

Candidate terms are proposed descriptors that have not gone through all acceptance procedures. These terms should be marked by a special symbol or phrase in the term record. As soon as a candidate term is approved as a descriptor, the symbol or phrase should be deleted.

In an online system in which the thesaurus is linked to a single database, candidate terms are generally not displayed to users; in thesauri that are not linked in this way, candidate terms may be displayed.

8.6.2. Provisional Terms

Where no published authority exists, descriptors for new concepts may be labeled provisional. Provisional terms are checked periodically to determine whether there is literary warrant for establishing them as descriptors. When provisional terms are upgraded to descriptors, the label "provisional" is removed.

8.7. Level of Specificity

The addition of highly specific descriptors is usually restricted to the core area of the subject field covered by a thesaurus, because the proliferation of such descriptors in fringe areas is likely to lead to a thesaurus that is difficult to manage (see section 5.2.4.1). Although the cost of computer storage of a very large thesaurus may be insignificant, the human cost of establishing relationships among numerous descriptors at the periphery of the domain is high.

In an organization that deals with documents covering more than one domain of knowledge, it may be necessary to develop a number of specialized thesauri, each linked to, and compatible with,

a general thesaurus that has a lower level of specificity, produced by the same organization. A model for this may be found in the *Macrothesaurus* of the United Nations. (See also section 10.12.2.)

8.8. Unassigned Descriptors

When hierarchies are being established in a thesaurus, descriptors that have not yet been used in indexing are frequently admitted into the thesaurus on the grounds that they are necessary to complete a hierarchy (for example, as broader terms), and that they have potential value as indexing terms in their own right, where the thesaurus is linked to a particular database. These descriptors should be marked by a special symbol in the hierarchical display and in the term record, or by a phrase such as "not yet used." (The phrase would not be necessary in cases where all thesaurus displays and term records indicate the number of postings.) As soon as the descriptor is used in indexing, the symbol or phrase should be deleted.

8.9. Announcement and Deposit of Published Thesauri

8.9.1. Notification of Intent

When an organization has decided to compile and publish a new thesaurus, it should submit a notice of intent to an appropriate professional journal, to minimize the possibility of two organizations simultaneously developing thesauri in the same domain.

8.9.2. Pilot Thesaurus

A new thesaurus should be tested by means of a series of trial runs prior to publication in printed or electronic form. A pilot version should be distributed to a select group of users, and suggestions for changes of terms and their relationships should be considered by the editors. The review of established thesauri is treated in section 9.1.

8.9.3. Deposit with a Clearinghouse

A copy of the first and any subsequent edition of a printed thesaurus or a printout of a thesaurus maintained only in electronic form should be deposited with the appropriate clearinghouse (see Appendix D).

9. Thesaurus Maintenance

This section suggests procedures for adding, modifying, and deleting terms. The term record should note the date of each change and identify the individual responsible for it.

9.1. Updating the Vocabulary

Thesauri are reflections of language, and they are therefore dynamic instruments. Policies and procedures should be established for periodic review of terminology, establishment of new descriptors, and replacement of obsolete descriptors, especially in fields where the terminology changes rapidly. Thesaurus editors should update the thesaurus at intervals that will be determined by the frequency and volume of changes made, and by the method of distribution. Printed revisions will necessarily be produced less frequently than revisions to a computerized system, in which updates can be accessible more quickly than is possible with print — instantaneously in some systems.

9.1.1. Addition of Descriptors

Whenever an appropriate descriptor or combination of descriptors cannot be found in a thesaurus, the indexer or searcher should nominate a new descriptor as a candidate term (see section 8.6.1). The nomination, including an adequate definition of the term, may be submitted electronically, possibly via an on-screen form, or by means of a printed form. A sample form is in Figure A19. New terms and those nominated for reconsideration should be reviewed by the editor of the thesaurus, and preferably by an editorial board that functions in the same manner as the group responsible for the original edition. The following decisions should be made for each term:

a) Should this term be added to the thesaurus, or is it already covered by an existing descriptor or combination of descriptors? If it is covered, should a USE reference be added?
b) Are scope notes or other notes needed? (See section 3.2.2.)
c) If the term is to be added, what is the correct form, considering syntax, number, and spelling? (See section 3.)
d) How should the new term be interrelated with existing descriptors on a hierarchical and related-term basis? (See section 5.)

9.1.2. Modification of Existing Descriptors

Indexers and searchers should also be able to propose modifications to existing descriptors or their relationships, explaining the rationale and supplying supporting documentation for the proposed changes. Like candidate term nominations, such proposals may be communicated electronically or via printed forms. (Figure A19, Thesaurus Term Review Form, accommodates both candidate terms and modifications to existing descriptors.)

Such proposed changes should be considered by the thesaurus editor and board, using the criteria for descriptor selection in sections 3 and 4. If a descriptor is modified, the date of the change should be recorded in the history note (see section 8.5), and a USE reference should be made from the old form to the new form.

If the thesaurus is used in an indexing system, the date on which an old descriptor was last assigned should be included in the history note. If the relationships are modified, a record of the old ones should be maintained in the history note as well.

9.1.3. Deletion of Terms

Overused descriptors and descriptors that have been assigned very infrequently in indexing should be considered candidates for modification or deletion, as both kinds of term are generally ineffective in retrieval. In some cases, an overused descriptor can be replaced by two or more descriptors of greater specificity, or these can be added to it as narrower terms.

9.2. Vocabulary Updates and Database Records

Whenever a descriptor is modified or deleted from a thesaurus, the impact of the change on the ability to search previously indexed database records must be considered, unless the modified or deleted descriptor has never been assigned. The problem of searching modified or deleted descriptors can be handled in several ways. The choice of method depends upon the size of the database, available human and financial resources, the nature and number of changes to be made, and whether automated procedures exist.

a) A deleted or modified descriptor may be retained in the thesaurus for retrieval purposes only. It should be marked, e.g., "for retrieval purposes only," and the date of its change in status should be recorded in the history note and displayed to users. If it is superseded by a newer descriptor, a cross-reference to the new descriptor should be provided. The superseded descriptor should also be listed in the history note of the newer descriptor, with the dates for which the former is valid.

b) One-for-one descriptor changes, where one term simply replaces another, can be implemented in the database automatically if resources permit. In very large systems, the cost of such file maintenance may be prohibitive. The history note in the term record then serves as a guide for retrospective searching.

c) More complicated term changes require human intervention, e.g., where two or more specific descriptors are added to replace a more general term or become narrower terms to it.

Deleting a descriptor in the database without replacing it with another descriptor may, over time, seriously reduce access to documents if no replacement indexing is carried out. If it is discovered, however, that a descriptor has been established and assigned in error, once the descriptor has been deleted from database records, its record may be deleted from the thesaurus as well.

9.3. History of Changes

Each descriptor requires a history note recording its date of entry (see section 8.5). The note should also track the history of modifications (see section 9.1.2), recording previous forms with reasons for the change. Where obsolete descriptors are retained in the thesaurus but are not assigned in indexing, the date of replacement and reasons for it should be given. Where terms are deleted completely, it may be desirable to maintain a separate file of deleted descriptors, retaining a record of their scope and history in order to track such decisions in case a term from this group is later reconsidered.

Examples of history notes of various types and with varying degrees of detail are found in the figures in Appendix A:

- Figure A1: The Code DA notes date of entry into the thesaurus.
- Figure A4: The entry:

  Action Programs (Community) (1966 1980)
  USE COMMUNITY ACTION

indicates the dates for which the former descriptor was valid.

- Figure A5 illustrates the opposite pattern in the last entry:

  FEDERAL REGION X
  (Prior to June 1982 this concept was indexed to PACIFIC NORTHWEST REGION.)

- Figure A14 documents a change in an entry term:

  FERTILIZATION IN VITRO
  Apr 79; TEST-TUBE FERTILIZATION was TEST TUBE BABIES see FERTILIZATION IN VITRO Apr-Dec 1979
  X TEST-TUBE FERTILIZATION

- Figure A14 also provides an example of the record of a change in status of a descriptor:

  FERULA
  was see under PLANTS, TOXIC 1986-90

This history note means that, for the years indicated, the descriptor was not authorized for use in

the printed index, only in the online database. The *see under* reference led the indexer or searcher to a broader descriptor that was authorized for use in the print index.

10. Thesaurus Management Systems

Because thesauri are generally used in the context of information storage and retrieval systems, the design or choice of a thesaurus management system is necessarily dependent on and should be an integral part of the overall design of the information system of which it is a component.

Maintainers of thesauri should have access to systems that can deal with the special needs of a thesaurus. It should not be necessary to adapt the thesaurus to an inadequate system. The system should be flexible enough to allow thesaurus managers to take advantage of emerging technologies.

The following are general recommendations for features of thesaurus management software to be used by thesaurus maintainers. See also the recommendations in section 7 regarding screen display for users.

10.1. Typography and Sorting

The system should allow printing or screen display in both capitals and lowercase letters, and it should be able to handle any special characters required by the subject field or language of the thesaurus. (See section 3.7, Capitalization, Punctuation, and Nonalphabetic Characters.) The sort sequence should treat the capital and lowercase versions of a letter as equivalent and should comply with the filing guidelines presented for printed thesauri in section 6.3.5.

10.2. Term Records and Display Format

At minimum, term records (see section 8.5) should contain the data and specify the relationships recommended in this standard; but an individual thesaurus may require additional data, such as term weights for retrieval, special relationships, or codes from standards such as the USMARC Authority Format. The system should be capable of handling hierarchical thesaurus displays, where required.

The system should also be able to distinguish between the nonpreferred terms (USE references) that should be displayed to thesaurus users and those that should not be displayed (e.g., close spelling variations and common typographical errors entered to enhance computerized retrieval).

10.3. System Constraints

There should be no limits on the number of characters in a term, the number of levels in a hierarchy, or the number of relationships (broader, narrower, associative, or equivalence) that can be established for a given descriptor.

10.4. Searching

Thesaurus maintainers, indexers, and intermediaries should be able to search for and retrieve complete or truncated terms, including nonpreferred terms. System sensitivity to capitals and lowercase letters may lead to retrieval failure; a search term in capitals should therefore match a lowercase descriptor or entry term and vice versa.

Browsing displays should be available and should include hierarchical, alphabetical, and (optionally) KWIC or KWOC displays. (The need for the last type of display may be obviated by a text-word search capability.) From any of these displays, it should be possible to view a term in the context of its relationships as well as a complete term record.

Another desirable feature, particularly in large systems, is to allow for retrieval of descriptors and entry terms on the basis of quantitative characteristics, such as the number of characters or words in a term. Such data are useful for the determination of requirements for field length in term records, for the calculation of system capacity, as well as for projections of growth of the thesaurus.

10.5. Editing

Maintainers should be able to create and edit terms and to navigate editing screens easily using templates or prompts. When new relationships are being entered, or old ones deleted, the existing relationships should be visible on the screen. It should be possible to edit a descriptor's position in a hierarchy, and to simultaneously display multiple hierarchies for a descriptor where these exist.

Editing capabilities should include standard word processing features, i.e., the ability to add, modify, and delete characters, words, and lines without rekeying an entire field, and to navigate easily backward and forward between fields and relationship types.

A sample menu from a thesaurus management system is shown in Figure A20.

10.6. Error Checking

The system must be able to detect duplicate terms, whether entry terms or descriptors. Duplicates detected by the system should be

displayed to the thesaurus editor, who will judge whether they are true duplicates or homographs requiring qualification. The system must be able to identify duplicates with typographic variations (capitals and lowercase letters) for qualification, e.g., **mercury** (metal) and **Mercury** (planet). It should be able to check for inappropriate coding and content, as well as errors in term relationships and inconsistencies in the modification or deletion of terms, as described in the following *sections*.

10.7. Cross-References

When cross-reference links are created between terms, the system should:

a) check for the existence of a term chosen as a broader, narrower, related, or USE term, and check for the validity of references, so that:

- no cross-references are made to nonexistent terms;
- the only type of reference that leads from a nonpreferred to a preferred term is a USE reference. Nonpreferred terms have no BT, NT, or RT references;
- a term is not related to itself;
- no conflicting references are made for the same descriptor pairs (e.g., both a BT/NT and RT/RT reference);
- duplicate cross-references are eliminated.

b) create reciprocal references for BT/NT (or BTG/NTG and BTP/NTP), RT/RT, and USE/UF pairs. If the reciprocal descriptor for a BT, NT, or RT reference does not exist, the system should inquire whether a new descriptor should be created, but should not create it automatically. However, if the reciprocal term of a UF reference does not exist, the system may create that term. See section 5.1.2 for logical formulations of the relationships.

Examples:

editor creates	**passenger cars** UF cars
system creates	cars USE **passenger cars**
editor creates	**motor vehicles** NT **passenger cars**
system creates	**passenger cars** BT **motor vehicles**

c) maintain these references reciprocally, so that when a term is modified or deleted, the references to it are changed accordingly in the records for all the terms related to it. For example, if the editor changes **passenger cars** to **automobiles** in the above example, the system should make the same change throughout the thesaurus, e.g., **motor vehicles** NT **automobiles**.

d) When a thesaurus is integrated with a hypermedia system, the appropriate standards for checking and maintaining hypermedia links apply.

10.8. Term Deletion

When a term is deleted, the system should first display the term and then prompt for verification of the deletion. The system must make deletions reciprocally so that all links to and from deleted terms are deleted with the term (see Figure A21). The system should inform the thesaurus editor when the deletion creates an orphan term (a descriptor without associative or hierarchical relationships).

10.9. Candidate Terms

The system should include a mechanism for the entry of candidate terms, designated as such, and candidate relationships, which are not admitted to the thesaurus until reviewed and approved by thesaurus editors. It should be possible to retrieve, display, and print out these candidate terms and relationships. Candidate relationships should be reciprocated automatically. Whoever enters the proposed term should be able to record its source and to enter suggested relationships and a proposed scope note, because that person may be looking at relevant explanatory material at the time of entry. The structure of candidate term records should be the same as that of descriptor records (see section 8.5), because the former may be upgraded to the latter.

It is also useful for the system to be able to capture terms used by database searchers, as these may be useful entry terms for the thesaurus. The system may also provide a mechanism for searchers to propose candidate terms directly.

A merge or transfer function that allows for the comparison and merger of two similar terms can be useful in choosing among candidate terms entered by different indexers, or where revisions of a thesaurus are necessary in a quickly changing environment. The preferred term receives the relationships and notes from the nonpreferred term, and the nonpreferred term becomes a USE reference.

10.10. Field Definition

The system should provide for the definition of fields other than those enumerated for term records

(section 8.5), for purposes such as assigning codes from other systems to descriptors, e.g., Standard Industrial Classification (SIC) codes to the names of industries; or providing term definitions as well as scope notes. (Scope notes often differentiate between two descriptors for indexing purposes; definitions give the dictionary meaning.)

It may also be useful to be able to define and name specialized relationship types which are more specific than USE/UF, and BT/NT or RT/RT, e.g., designating abbreviations, or expressing the exact nature of various types of narrower terms, such as the subsidiary of a company. See section 5 for a discussion of term relationships, and Figure A22 for an example of a screen display featuring specialized codes.

10.11. Reports

It should be possible to obtain printouts and screen displays in whatever format the user requirements for an individual thesaurus dictate. (See section 6 on print displays and section 7 on screen displays.) In addition to reports designed for both users and maintainers, thesaurus managers need specific "housekeeping" reports. These reports may be displayed on a screen or printed, but in either case, they should be clear and legible and should use space and paper as economically as possible. Section 6 discusses the space implications of various printed thesaurus display formats. Screen displays should follow the guidelines in section 7. Thesaurus reports may include, but are not limited to, the following:

a) term listings, including:
   - terms added or modified since a certain date;
   - descriptors that do or do not have a particular type of note or relationship (SN, UF, BT, NT, or RT);
   - orphan terms, with no hierarchical or associative relationships;
   - top terms;

b) displays of term relationships, including:
   - all the relationships for an individual term;
   - hierarchical displays;
   - listings at various levels of a hierarchy;
   - reports of special links created for a given thesaurus;
   - the entire hierarchical context for an individual term;

c) statistical reports on terms and their characteristics, including:
   - the number of descriptors and of all terms in the thesaurus;
   - the number of terms featuring a certain characteristic, for example, those that consist of three words or those that end in a certain string of characters, e.g., *itis*;
   - the number of terms added, modified, deleted, or merged since a certain date;
   - the number of BTs, NTs, UFs, RTs, SNs, or other value in the thesaurus;
   - the average number of characters per term and the size of the longest term;
   - postings reports for thesauri linked to databases.

10.12. Subsets

10.12.1. Copying Function

It is useful to be able to copy a subset of terms from one thesaurus to another where more than one thesaurus is being maintained on a system.

10.12.2. Microthesauri

Omnibus indexing services (organizations that produce multiple specialized indexes through a single indexing operation) or documentation centers may maintain a thesaurus of broad scope, subsets of which, known as microthesauri, are used in specific indexing products. When such subsets are extracted, only relationships to descriptors in the microthesaurus should be included. A microthesaurus must be able to stand alone.

Microthesauri may contain additional highly specific descriptors required for their application, but such descriptors should map to the hierarchical structure of the broad thesaurus.

References

This list includes sources cited in the text as well as in the Glossary, which follows. References to older thesaurus standards that are not cited in the text are found at the end of the Foreword. A bibliography of manuals of thesaurus construction is in Appendix E.

[AACR2] Anglo-American Cataloguing Rules, 2nd ed. Rev. Michael Gorman and Paul W. Winkler, Eds. Chicago: American Library Association, 1988.

ALA Filing Rules. Chicago: American Library Association, 1980.

ALA-LC Romanization Tables. Washington, DC: Library of Congress, 1991.

[ANSI Z39.4] American National Standard for Library and Information Sciences and Related Publishing Practices - Basic Criteria for Indexes. New York: American National Standards Institute, 1984.

Art and Architecture Thesaurus. Second edition. Toni Petersen, Director. New York: Oxford University Press, 1994. 5 vols.

British Standard Alphabetical Arrangement and the Filing Order of Numerals and Symbols. London: British Standards Institution, 1985. (BS 1749:1985).

[ERIC] Thesaurus of ERIC Descriptors. 12th edition. James E. Houston, Ed. Phoenix, AZ: Oryx Press, 1990.

Inspec Thesaurus, 1993. London: Institution of Electrical Engineers, 1993.

International Energy Subject Thesaurus. Oak Ridge, TN: U.S. Department of Energy, Office of Scientific and Technical Information, 1990. (ETDE/PUB-2; DE90008750).

[ISO 2788] Documentation—Guidelines for the Establishment and Development of Monolingual Thesauri, 2nd ed. Geneva: International Organization for Standardization, 1986.

[ISO 5127/1] Information and Documentation - Vocabulary - Part 1. Basic Concepts. Geneva: International Organization for Standardization, 1983.

[ISO 5127/6] Information and Documentation - Vocabulary - Part 6. Documentary Languages. Geneva: International Organization for Standardization, 1983.

[ISO 5964] Documentation—Guidelines for the Establishment and Development of Multilingual Thesauri. Geneva: International Organization for Standardization, 1985.

Library of Congress Filing Rules. Prepared by John C. Rather and Susan C. Biebel. Washington, D.C.: The Library, 1980.

Macrothesaurus for Information Processing in the Field of Economic and Social Development. 3rd ed., prepared by Jean Viet. New York: United Nations, 1985.

[MeSH, 1992a] Medical Subject Headings: Annotated Alphabetic List, 1992. Bethesda, MD: National Library of Medicine, 1991. (NTIS: PB 92-100031).

[MeSH, 1992b] Medical Subject Headings: Supplement to *Index Medicus* vol. 33, 1992. Bethesda, MD: National Library of Medicine, 1992. (NIH Publication No. 92-265).

[MeSH, 1994c] Medical Subject Headings: Tree Structures, 1994. Bethesda, MD: National Library of Medicine, 1993. (NTIS: PB 94-100054).

NASA Thesaurus. NASA, Scientific and Technical Information Branch. Springfield, VA: NTIS, 1988. 3 vols. Vol. 1: Hierarchical Listing. Vol. 2: Access Vocabulary. Vol. 3: Definitions.

[TEST] Thesaurus of Engineering and Scientific Terms. New York: Engineers Joint Council, 1967.

USMARC Format for Authority Data, Including Guidelines for Content Designation. Network Development and MARC Standards Office. Washington, D.C.: Cataloging Distribution Service, Library of Congress, 1987.

Glossary of Terms

The following definitions are derived from several related standards and specialized dictionaries, but all have been modified for the purposes of this standard. References to other standards appear in brackets. Underlined terms within definitions are defined elsewhere in this glossary.

assignment indexing. An indexing method by which descriptors or subject headings from a controlled vocabulary are selected by a human or computer to represent the concepts or items in a document. The terms may or may not occur in the document. Cf. **derivative indexing.**

authority file. A set of records of established headings and the cross-references to be made to and from each heading, often citing the authority for the preferred form or variants. Types of authority files include name authority files, subject authority files, and thesauri.

autoposting. *See* **up-posting.**

bound term. A descriptor consisting of a compound term. (The term was originated by Mortimer Taube in his *Studies in Coordinate Indexing*, vol. 1, 1953, p. 43.)

broader term. A descriptor to which another descriptor or multiple descriptors are subordinate in a hierarchy. The relationship indicator for this type of term is BT.

candidate term. A term considered for admission into a thesaurus because of its potential usefulness as a descriptor. Cf. **provisional term.**

category. A grouping of descriptors that are semantically or statistically associated, but which do not constitute a strict hierarchy based on genus-species or part-whole relationships.

classified display. *See* **tree structure.**

compound term. A term consisting of more than one word. A compound term may be selected as a descriptor.

concept. A unit of thought, formed by mentally combining some or all of the characteristics of a concrete or abstract, real or imaginary object. Concepts exist in the mind as abstract entities independent of terms used to express them.

controlled vocabulary. A subset of the lexicon of a natural language. A list of preferred and nonpreferred terms produced by the process of vocabulary control. Types of controlled vocabulary include subject heading lists and thesauri.

cross-reference. A direction from one term to another. Cross-references are of three types: (a) *equivalence relationship*, beginning with the word *see* or USE, which leads to one or more descriptors that are to be used instead of the term from which the cross-reference is made; (b) *associative relationship*, beginning with the words *see also*, or related term (RT), which leads from one descriptor to other descriptors that are related to or associated with it in the context of a thesaurus; (c) *hierarchical*, beginning with the words broader term (BT), or narrower term (NT), which represent generic and specific relationships, respectively. (It is possible to display hierarchical relationships without the use of cross-references.)

derivative indexing. An indexing method by which words occurring in the title or text of a document are extracted by a human or computer to serve as indexing terms.

descriptor. A type of heading that is a term chosen as the preferred expression of a concept in a thesaurus.

difference. In a compound term, one or more components that serve to narrow the extension of a focus and specify one of its subclasses. Also known as 'modifier.'

document. Any item, printed or otherwise, that is amenable to cataloging and indexing. The term applies not only to written and printed materials in paper or microform versions (e.g., books, journals, maps, diagrams), but also to nonprint media (e.g., machine-readable records, transparencies, audiotapes, videotapes) and, by extension, to three-dimensional objects or realia (e.g., museum objects and specimens).

entry term. The nonpreferred term in a cross-reference that leads to a descriptor in a thesaurus. Also known as 'lead-in term.' The relationship indicator for this type of term is USE; its reciprocal is UF (used for).

entry vocabulary. The set of nonpreferred terms (USE references) which lead to descriptors in a thesaurus. (This term is used by some thesaurus designers to represent the preferred as well as the nonpreferred terms in a thesaurus.)

eponym. A term incorporating the name of a real or mythical person, generally the discoverer of a phenomenon or inventor of an object, e.g., Herculean labor, Parkinson's disease, pasteurization.

facet indicator. *See* **node label.**

false drop. An irrelevant reference retrieved when natural language terms or descriptors are postcoordinated.

flat format. An alphabetical display format of thesauri in which only one level of broader terms and one level of narrower terms are shown for each descriptor.

focus. In a compound term, the noun component that identifies the class of concepts to which the term as a whole refers. *See also* **difference.**

free text. Antonym of controlled vocabulary. Natural language terms appearing in documents, which may complement descriptors in an information storage and retrieval system. In free text searching, descriptors may also be retrieved. Cf. **keyword.**

generic posting. (1) In thesauri, the treatment of narrower terms as quasi-synonyms, e.g., **furniture** UF beds; UF sofas. (2) In indexing and subject cataloging, the assignment of a broader term instead of a specific term, e.g., **furniture** to a document on sofas. Cf. **up-posting.**

generic structure. A thesaurus format that indicates all hierarchical levels of descriptors within an alphabetic display by means of codes, indention, and/or punctuation marks.

gloss. *See* **qualifier.**

heading. A preferred name or term. Types of heading include proper name headings (which may be called identifiers), subject headings, and descriptors. A heading may include a qualifier.

hierarchy. Generic (broader)-specific (narrower) or whole-part relationships, which are generally indicated in a thesaurus through codes or indention. *See also* **broader term; narrower term.**

history note. A note in a term record in a thesaurus that provides the date of entry of a descriptor as well as the history of modifications to its scope, relationships, etc.

homograph. One of two or more words that have the same spelling, but different meanings and origins. In thesauri, homographs are generally distinguished by qualifiers.

identifier. (1) A proper name (or its abbreviation or acronym) of an institution, person, place, object, or process, optionally treated as a category of heading distinct from descriptor. Identifiers may be held in a separate file (cf. authority file), and their form may be controlled (e.g., the name of an international organization having different names in various languages, only one of which is selected). (2) In some systems, a provisional term that may be upgraded to descriptor status, or a highly specific term that is not eligible for descriptor status, but which is considered useful for retrieval and is assigned to one or more documents without vocabulary control.

indexing. An operation intended to represent the results of the content analysis of a document by means of a controlled indexing language or by natural language. [ISO 5127/1]

indexing language. A controlled vocabulary or classification system and the rules for its application. An indexing language is used for the representation of concepts dealt with in documents and for the retrieval of such documents from an information storage and retrieval system. [ISO 5127/1]

indexing term. The representation of a concept in an indexing language, generally in the form of a noun or noun phrase. Descriptors, subject headings, and heading-subheading combinations are examples of indexing terms.

information storage and retrieval system. A set of operations and the associated equipment, software, and documentation by which documents are indexed and the records are stored, so that selected records can be retrieved in response to requests employing commands that can be handled by the system.

keyword. A word occurring in the natural language of a document that is considered significant for indexing and retrieval. Cf. **free text.**

KWIC (Key Word In Context) **index.** A type of index, arranged alphabetically, in which each significant word in a string of text serves as an access point, by being graphically emphasized and surrounded by the rest of the string. The keyword is generally in a centered column and is followed on the right by the continuation of the string, which provides the context. The balance of the string, if any, is positioned to the left of the keyword. A KWIC index of descriptors and entry vocabulary may serve as an auxiliary alphabetic display in a thesaurus.

KWOC (Key Word Out of Context) **index.** A type of index, arranged alphabetically, in which each significant word in a string of text serves as an access point, usually positioned in the left-hand column of a page, followed by the complete string. The keyword may therefore not be in the immediate context of the words that surround it. A KWOC index of descriptors and entry vocabulary may serve as an auxiliary alphabetic display in a thesaurus.

lead-in term. *See* **entry term.**

lexeme. A lexical unit in a language; a vocabulary item, which is not necessarily limited to a single word.

literary warrant. Justification for the representation of a concept in an indexing language or for the selection of a preferred term because of its

frequent occurrence in the literature. Cf. **user warrant.**

microthesaurus. A subset of a thesaurus, covering a limited range of topics within the domain of the thesaurus. A microthesaurus may contain highly specialized descriptors that are not in the broad thesaurus. Such descriptors should map to the hierarchical structure of the broad thesaurus. A microthesaurus is internally consistent with respect to relationships among terms.

modifier. *See* **difference.**

narrower term. A descriptor that is subordinate to another descriptor or to multiple descriptors in a hierarchy. The relationship indicator for this type of term is NT.

natural language. A language used by human beings for verbal communication. Words extracted from natural language texts for indexing purposes without vocabulary control are often called keywords.

near-synonym. *See* **quasi-synonym.**

node label. A "dummy" term, often a phrase, that is not assigned to documents when indexing, but which is inserted into the hierarchical section of some thesauri to indicate the logical basis on which a class has been divided; also known as a 'facet indicator' [ISO 2788]. Node labels may also be used to group categories of related terms in the alphabetic section of a thesaurus.

nonpreferred term. One of two or more synonyms or lexical variants that serves as an entry term.

orphan term. A descriptor that has no associative or hierarchical relationship to any other descriptor in a thesaurus.

polyseme. A word with multiple meanings. In spoken language, polysemes are called homonyms; in written language they are called homographs. Only the latter are relevant to thesauri designed for textual information.

postcoordination. The combining of descriptors at the searching stage rather than at the subject heading list construction stage or indexing stage.

postings. The number of documents to which a descriptor is assigned.

precoordination. The formulation of a multiword subject heading or the linking of a heading with a subheading at the subject heading list construction stage or at the indexing stage to express a compound concept, e.g., **cataloging of serials, cataloging—serials,** or **serials cataloging.** Precoordination differs from the establishment of compound terms as descriptors.

preferred term. One of two or more synonyms or lexical variants selected as a descriptor.

provisional term. A descriptor with temporary status in a thesaurus; it often represents a new concept in a field in which the terminology has not yet been standardized. Cf. **candidate term; identifier** (definition 2).

qualifier. A defining term, also known as 'gloss,' used in a controlled vocabulary to distinguish homographs. A qualifier is considered part of a descriptor, subject heading, or entry term, but is separated from it by punctuation (the qualifier is generally enclosed in parentheses).

quasi-synonym. A term whose meaning is not exactly synonymous with that of another term, yet which may nevertheless be treated as its equivalent in a thesaurus.

related term. A descriptor that is associatively but not hierarchically linked to another descriptor in a thesaurus. The relationship indicator for this type of descriptor is RT.

relationship indicator. A word, phrase, abbreviation, or symbol identifying a semantic relationship between terms. Examples of relationship indicators are USE, UF (used for), and RT (related term).

romanization. The conversion of a non-roman script by means of transcription or transliteration or a combination of the two methods.

scope note. A note following a descriptor explaining its coverage, specialized usage, or rules for assigning it.

sibling. A descriptor that shares the same broader term (one level higher) as other descriptors.

stop list. A list of words considered to be of no value for retrieval. It consists primarily of function words—articles, conjunctions, and prepositions—but may also include words that occur very frequently in the literature of a domain.

subheading. A term appended to a heading in order to modify or delimit the heading by indicating a particular aspect or relationship pertaining to it. A term with a subheading may be subject to further modification. Cf. **precoordination.**

subject heading. A word or phrase, or any combination of words, phrases, and modifiers used to indicate the overall content of a document or part of a document. Precoordination of terms for multiple and related concepts is a characteristic of subject headings that distinguishes them from thesaurus descriptors.

subject heading list. An alphabetical list of subject headings with cross-references from non-

preferred terms and links to related terms. These lists often include separate sequences of standardized subheadings that may be combined with all or only some subject headings. Rules for applying subheadings usually accompany such lists.

synonym. A word or term having exactly or very nearly the same meaning as another word or term. [Cf. ISO 5127/6]

systematic display. *See* **tree structure.**

term. One or more words designating a concept. *See also* **compound term; descriptor; entry term.**

term record. A set of fields of information about a descriptor in a thesaurus, including the history of the descriptor, its relationships to other terms, and, optionally, authorities for the descriptor.

thesaurus (pl. thesauri, also thesauruses). A controlled vocabulary arranged in a known order in which equivalence, homographic, hierarchical, and associative relationships among terms are clearly displayed and identified by standardized relationship indicators, which must be employed reciprocally. Its purposes are to promote consistency in the indexing of documents, predominantly for postcoordinated information storage and retrieval systems, and to facilitate searching by linking entry terms with descriptors. Thesauri may also facilitate the retrieval of documents in free text searching.

Note: The term "thesaurus" is the Latin form of the Greek word *thesauros*, originally meaning "treasure store." In the 16th century, it began to be used as a synonym for "dictionary" (a treasure store of words), but later it fell into disuse. Peter Mark Roget resurrected the term in 1852 for the title of his dictionary of synonyms. The purpose of that work is to give the user a choice among similar terms when the one first thought of does not quite seem to fit.

A hundred years later, in the early 1950s, the word "thesaurus" began to be employed again as the name for a word list, but one with the exactly opposite aim: to prescribe the use of only one term (a "descriptor") for a concept that may have synonyms. A similarity between *Roget's Thesaurus* and thesauri for indexing and information retrieval is that both list terms that are related hierarchically or associatively to descriptors, in addition to synonyms.

top term. The broadest descriptor in a thesaurus hierarchy, sometimes indicated by the abbreviation TT.

transcription. The process of recording the phonological and/or morphological elements of a language in terms of a specific writing system.

transliteration. The process of recording the graphic symbols of one writing system in terms of the corresponding graphic symbols of another writing system.

tree structure. A thesaurus display format in which the complete hierarchy of descriptors is shown. Each descriptor is assigned a tree number or line number which leads from the alphabetical display to the hierarchical one; the latter is sometimes called 'systematic display' or 'classified display.'

up-posting. The automatic assignment of broader terms in addition to the specific descriptor by which a document is indexed. Also known as 'autoposting.' Cf. **generic posting.**

user warrant. Justification for the representation of a concept in an indexing language or for the selection of a preferred term because of frequent requests for information on the concept or free-text searches on the term by users of an information storage and retrieval system.

vocabulary control. The process of organizing a list of terms (a) to indicate which of two or more synonymous terms is authorized for use; (b) to distinguish between homographs; and (c) to indicate hierarchical and associative relationships among terms in the context of a thesaurus or subject heading list. *See also* **controlled vocabulary.**

Enriched thesauri
and their uses in information retrieval and storage

Discussion paper

Michiel Hazewinkel
CWI, Amsterdam
POBox 94079, 1090GB Amsterdam
mich@cwi.nl

Introduction.

The subject of this discussion paper is information storage and, especially, information retrieval from large and very large collections of objects. The focus is on scientific objects such as papers, tables, programs, handbooks, manuals, ... All of these will be referred to here and below as documents. They can be quite heterogeneous in form and content and the storage medium can be varied, though it is assumed that at least adequate metadata (see below for this concept) are, or will be, attached in machine readable form.

Quite apart from the sheer size of the problem there are a variety of reasons (for instance linguistic ones having to do with morphological variations, synonyms, and homonyms) that indicate—personally I would put it much stronger—that full text search is not a real alternative. This has always been the major reason to work with a "controlled vocabulary", here interpreted as a standard list of key phrases for a given section of science and/or technology, [2].

Another main reason is multilinguality. It is feasible to have essentially the same thesaurus (of key PHRASES (possibly with additional identifiers to deal with ambiguities)) in several languages with good explicit correspondences. It is in any case far simpler to realize such a thing than to do anything like automatic translation.

2. Metadata. The key to dealing with (very) large collections of scientific documents is metadata. This concept includes such things as bibliographic data such as CIP data and Library of Congress cataloguing data; it also includes classifications in terms of one of more (standard) classification schemes and attaching a number of key phrases to a document. The importance of metadata is illustrated, e.g., by the very substantial effort that Elsevier puts continuously into maintaining the thesaurus behind EMBASE, the database of Excerpta Medica, [4], and the even more considerable effort involved in adding adequate metadata in the form of key phrases and words from that thesaurus to each document of EMBASE/Excerpta Medica.

The question of metadata rather naturally divides into two interrelated parts:

- Questions of design and implementation of a framework, or container, for metadata, including how to translate one metadata scheme to another.
- Questions concerning what should be put into the various slots of the container, including how these data should be used for information retrieval purposes and how to generate metadata for a given document (semi-)automatically.

For instance the Dublin Core famework (named for Dublin, Ohio, rather than Dublin, Ireland) consists of the elements: Subject, Title, Author, Publisher, Other agent, Date, Object type, Form, Identifier, Relation, Source, Language, Coverage. Most of these elements can be replicated as often as desired or necessary, [10]. Further work, extending the Dublin core, on developing containers for Metadata is in progress under the name Warwick framework , [5]. This is led by NCSRTL, has the participation of Library of Congress and the NSF/ARPA/NASA digital library

initiative and involves the national libraries of the UK, Norway, Finland, The Netherlands, Australia, New Zealand, and OCLC, NCSA, etc.

For most of these elements what kind of information should be put into them is fairly clear. A clear exception is the element *subject*, which encompasses such things as classifications according to various classification schemes and key phrases describing the subject matter of the document. In addition to the standardly used bibliographic data such as title and author (or identifier) this is the main source of data for information and document retrieval at the metadata level.

Much of the what follows below is adressed specifically to the matter of searching in quite large and very large collections. Whether these collections are distributed over a network or are all in the same place is largely irrelevant for the issues addressed here. Experienced users of the various search engines on the Web (WWW) will know how difficult it is to keep down the number of hits (while keeping them relevant) without sophisticated search tools. The problem is widespread and generally recognized.

3. Enriched thesaurus. Definition.

An enriched thesaurus for a given domain in science and/or technology consists of several components.

(i) The main and central component: a list L of key phrases sufficiently rich to describe the domain in question adequately. For instance for mathematics I estimate that L should have about 120 000 entries. For computer science about 90 000 should be adequate.The Encyclopaedia of Mathematics, [6], has about 30 000 entries in its index volume (not counting inversions and linguistic morphological variations). The thesaurus behind EMBASE and Excerpta Medica works with about 25 000 basic entries.

The key phrases in L should be long enough to carry real information (single words in science only rarely do so). All entries in L are prepositional noun phrases (PNP's). A prepositional noun phrase consists of noun phrases (NP's) (strings of adjectives and nouns with or without particles, linked together, possibly iteratively, by prepositions (such as 'of', 'for', 'on'). For example: "differential invariants of four dimensional manifolds".

(ii) A distance function on L which reflects how related items in L are and how likely they are to occur together in the same documents.

(Technical note. Storing a distance function on a set of this size explicitely, takes a horrible amount of storage space; fortunately there are far more efficient ways of doing this, [7].)

(iii) A partial order on the set L, indicating which phrases (concepts) are more general (broader terms) and which are less general (narrower terms).

(iv) A labeling of each item of L with a node (most often a leaf) of one or more classification schemes for the domain of inquiry in question, such as the PAC classification scheme for Physics and Astronomy, the INSPEC classification scheme for electrical engineering, computers and control, the MSCS (Mathematics Subject Classification Scheme) developed by the American Mathematical Society and FIZ/STN (Zentralblatt für Mathematik und Grenzgebiete).

Components (ii) and (iii), respectively, refine and replace the classical thesaurus notions of 'related terms' and 'more general term (broader term)', 'less general term (narrower term)'.

Note that the classification schemes used in component four, which are also part of the enriched thesaurus, also get enriched themselves in that each node acquires a list of key phrases carrying that node as classfication thus giving that node content and meaning far beyond the short description given by its name. Such descriptions are a valuable tool for researchers and others when assigning classifications to a document.

4.. Some background: controlled lists and error correcting codes.

The idea of using only part of all objects of a given kind, usually a quite small part (relative to the total) but a quite well distributed one, for purposes of description, understanding, manipulation, ... is one that occurs with frequency in various parts of science and technology. Thus one has for example (truncated) singular value decomposition in mathematics and signal processing, principal component analysis in econometrics (economics), factor analysis in psychology. These are mathematically basically the same and can also be used for intelligent information retrieval, [3]. Another instance is lexical stemming in linguistics and still another is coding theory in communication and information theory. In all cases the idea is to systematically get rid of such things as noise or contamination, unwanted variations, less important or unimportant factors.

The idea of a standard (but dynamically evolving) thesaurus is perhaps best thought of as a linguistic variant (and information retrieval variant) of an error correcting code.

5. Semi-automatic generation of an enriched thesaurus.

Take a large enough corpus and divide it into suitable chunks called documents. For instance take the 700 000 abstracts of articles in the STN/FIZ database Math (ZMG data), or take as documents the sections or pages of a large handbook or encyclopaedia such as the Handbook of Theoretical Computer Science, [9], or the Encyclopaedia of Mathematics, [6]. Now use a parser for prepositional noun phrases (PNP's) (or an automaton recognizing PNP's) or a software indexing program, to generate from these documents a list of key phrases, keeping track of what phrases come from what document. This gives component (i) of an enriched thesaurus and, using e.g. Hamming distance, the incidence matrix between documents and key phrases thus created also provides component (ii).

(Technical notes. For the results of a first exercise in this direction see [1, 8]. Such automata and/or indexing software packages exist in several varieties. The Hamming distance between two terms is the number of documents which contain the one term and not the other; more sophisticated variants, embodying the idea of weights (reflecting importance, length, ...) are obvious.)

To generate component (iii) (less essential in the present context and in view of the uses to be made of the enriched thesaurus; see below) one can generate the network of balls defined by the the discrete metric space of phrases defined by components (i) and (ii), [7].

(Technical note: obviously components (iii) and (iv) of an enriched thesaurus should be compatible in a suitable way; this can be done by applying a suitable clustering method to the discrete metric space of key phrases (thus yielding a bottomup classfication scheme). The general problem of finding a best partial order compatible with one or more classification schemes is largely open.)

Generating component (iv) of an enriched thesaurus is relatively staightforward if one works with a corpus like the ZMG data or a similar database maintained by STN/FIZ for theoretical computer science (to which adequate access has been arranged). The documents in these corpora are labelled with classification codes from the corresponding classification scheme. Similar access to the ACM and INSPEC (IEEE) databases will be rather necessary to do a good job for all of computer science.

5. Thesauri rather than one thesaurus. The title of this essay speaks of thesauri rather than one thesaurus. In my opinion, it is flatly impossible to generate all at once an adequate thesaurus for all of science or even of all of a largish field like physics. The field of mathematics is about as large as can be handled at once I think. The solution is to have several thesauri which may overlap and be of different levels of detail. Much like in a geographical atlas one has maps and charts which overlap and can have different scales (and be of different kinds).

(Technical note. There are intersting and open mathematical and algorithmic problems in obtaining maximally good overlap correspondences.)

6. Searching. Once one has an enriched thesaurus, in the sense defined above, in place there become available advanced intelligent search possibilities based on the contents of the subject fields, classification fields, title field, key word fields, and abstract field in the metadata container attached to each document in the collection at hand. Note that in the Dublin core design for metadata all these fields, except the title field, are replicas of "subject". These search possibilties can be seen as being of two kinds:

Direct thesaurus mediated search

Dialog mediated search.

Both come with a 'local search' refinement.

For the first kind the searcher types in a Boolean combination of items in the thesaurus list, the author list, the classification scheme lists, and obtains as response the number of hits. The answers thus obtained can be Booleanly combined with the results of other queries and the final result can be displayed as a list of documents.

For the second type of search the searcher types an arbitrary potential key phrase (or several). A *PNP comparer* answers with a list of the n closest phrases which actually occur in the thesaurus and asks the searcher which ones he would like to use in some Boolean combination. Here n is a number that can be specified by the searcher with a default value of for example 5.

(Technical note. A PNP comparer in the present context is very much like a decoder in coding theory. The general problem of defining an adequate distance on sentences for a natural language grammar which allows for the automatic correction of small grammatical mistakes is a long standing and much neglected open problem; the same problem is at hand here but only for the very simply structured class of PNP's and several admittedly ad hoc solutions suggest themselves.)

For the local search refinement remember that the space of key phrases comes with a metric (and so do the spaces of classification scheme items). Thus the searcher can specify a centre, which is a set of such items, and request a seach for specified other items which are within a searcher determined distance of the given centre.

For a final search possibility note that the documents in the total distributed repository have key phrases and classifications attached to them. The resulting incidence matrix also defines a metric on the space of documents in the total repository. Thus it is possible to specify a document centred search which can be of one of the above types or can even be full text if so desired.

7. Updating issues.

Science fields evolve. To remain useful an enriched thesaurus must be regularly updated. Also, certainly, besides using key phrases from the existing thesaurus, authors must be free to invent ones of their own and put them into the subject fields of the metadata container.

(Technical note. Should the key phrases in the subject fields in the Dublin core be marked depending on whether they come from a specified thesaurus or were freely invented?)

Dynamic updating of the classification schemes used ususally, of course, will be outside the scope of a group engaged in constructing a thesaurus for a given field. With this restriction, updating an enriched thesaurus as defined above is a relatively straightforward matter. It does require, however, a small *new key phrase detector*, which lists the new key phrases invented by the authors which contribute documents to the total repository. These become candidate new key phrases which can be incorporated, if found suitable, at some future time.

8. Multilinguality.

As suggested above automatic translation of a query in one language to one in another one is difficult to realize directly. However, it is possible, given adequate corpora, to do the 'thesaurus' generating job described above for several different languages. The resulting metric spaces

should be more or less similar and that can be used to establish reliable correspondences between them. More precisely, using existing dictionaries, the metric space structures can be used to provide suggestions for translations and to provide checks on the adequacy of the translations. Again it needs to be stressed that it is translations of KEY PHRASES rather than individual words which are needed

9. Atlas of science and technology.

As said above, I do not think that a job such as was described above can be done all at once for all of science, not even for all of a largish established and well defined discipline like physics. The aim is to generate thesauri (in a similar way) for many partially overlapping parts of science, and, where possible, to do it at varying levels of detail. Thus there will be many thesauri in varying levels of detail for varying parts of science and addressing different kinds of information. The whole can be likened to a geographical atlas with its many different charts at different scales and of different kinds (geologica, demographical, linguistic, mineralogical, political, ...). Just as in the case of such an atlas the correspondences (which in the case of a geographical atlas are rather obvious) need to be established. Here again the metric structures of the specific thesauri which make up the atlas, can play a guiding role.

10. Concluding remark.

The above is not the description of an existing thing. It is instead an outline of a possible project, that, in my opinion, can be carried out with basically existing tools, and that should be carried out.

References.

1. *Subject index volumes 101 - 150*, Theoretical Computer Science **150**:2 (1995), 197-313.
2. Jean Aitchison, Alan Gilchrist, *Thesaurus construction*, Aslib, 2-nd Edition, 1990.
3. Michael W Berry, Susan T Dumais, Gavin W O'Brien, *Using linear algebra for intelligent information retrieval*, SIAM Review **37**:4 (1995), 573-595.
4. Ian Crowlesmith, *Creating a treasure trove of words*, Elsevier Science World. 14-15, 1993.
5. Juha Hakala, Ole Husby, Traugott Koch, *Warwick framework and Dublin core set provide a comprehensive infrastructure for network resource description*, html document, 1996. URL: http://www.bibsys.no/warwick.html
6. M Hazewinkel (ed.), *Encyclopaedia of mathematics; 10 volumes*, KAP, 1988-1994.
7. Michiel Hazewinkel, *Classification in mathematics, discrete metric spaces, and approximation by trees*, Nieuw Archief voor Wiskunde **13** (1995), 325-361.
8. M Hazewinkel, *Preface to the index for volumes 101-150 of TCS*, Theoretical Computer Science **150**:2 (1995), 195-197.
9. Jan van Leeuwen (ed.), *Handbook of theoretical computer science*, Elsevier, 1990.
10. Stuart Weibel, Jean Godby, Eric Miller, *OCLC/NCSA metadata workshop report*, OCLC, 1995. URL: http://www.oclc.org:5046/oclc/research/conferences/metadata/dublin_core_report.

Saturday, 4 May 1966, Revised: Wednesday, 05 February 1997

STORIES

D-Lib Magazine
November 1998

ISSN 1082-9873

WWW -- Wealth, Weariness or Waste

Controlled vocabulary and thesauri in support of online *information* access

David Batty
CDB Enterprises, Inc.
Adjunct Professor, Graduate School of Library and Information Science
The Catholic University of America
davidbatty@aol.com

Abstract

This article offers some thoughts on the problems of access to information in a machine-sensible environment, and the potential of modern library techniques to help in solving them. It explains how authors and publishers can make information more accessible by providing indexing information that uses controlled vocabulary, terms from a thesaurus, or other linguistic assistance to searchers and readers.

Introduction

Effective communication in any context is made easier by the use of a common language that both parties understand. In human speech, common language may occasionally be ambiguous or misleading because of the way it is used, but usually, the quantity and feedback of exchanged messages allows the parties to communicate with each other. In the world of recorded information, there is not the same opportunity for immediate interaction, feedback and realization. Writers are not consistent in their use of all the words that might describe or refer to topics, so a searcher who chooses to scan the full text of a document must include in a search statement all the synonyms, related terms, and all levels of detail that the author might have used. The searcher is not talking to the author directly but to the document itself, a static instrument. Therefore we need a substitute for the result of the human, spoken interchange: the "Ah, yes, I see. What you mean is ... " That needed substitute is the controlled index language (thesaurus) that the indexer uses to interpret and represent the themes, concepts, and language of the author and that the searcher uses to interpret and represent sometimes vague expressions of a need to know.

Practical Experience

Unfortunately, in the USA the value of controlled index languages was ignored for many years. Many early online database providers were dissuaded from using a thesaurus for what seemed to be valid economic considerations: fear of the apparent expense of designing and using a thesaurus; and doubt that the size of the database would justify that expense. These apprehensions turned out to be ill-founded. Databases grew to almost unmanageable size; the cost of searching titles and abstracts far exceeded what would have been the investment in the development and use of a controlled index language.

With CD/ROM publication we have at our fingertips physical packages that can contain 650MB in a single unit -- roughly the equivalent of 200,000 digitally stored single-spaced, 8.5"x11" pages or 10,000 printed book pages stored as graphic images. But this accessibility is misleading, since we cannot flip through the CD/ROM anything like as easily as we can through a printed book or document file. And with the rapid growth in publication on the Web, we have available almost unlimited information resources -- even bearing in mind Sturgeon's Law: nine-tenths of anything is junk. The problem is to find information to meet a need without having to read all of it. Underlying this problem is editorial supervision. Even acknowledging the low cost of CD/ROM production, most publishers exercise care for the intellectual product, as do conventional book publishers, as do publishers on the Web. But when anyone can set up an accessible site and load into it anything an untidy mind chooses, the level of control is minimal. Full text searching will always be valuable for browsing in any size of file; but in large files, controlled language access searching will always support efficient retrieval.

This is something that even law firms, long convinced of the efficacy of whole text searching, are beginning to appreciate. Several studies of litigation support activities have shown that the use of whole text searching is not as effective as the use of even a simply designed controlled index language. For example, see "An evaluation of retrieval effectiveness for a full-text document-retrieval system" by David C. Blair and M. E. Maron and follow-up articles by Gerard Salton, and Blair and Maron. Litigation support companies hired by law firms to gather discovery documents and have them ready for attorneys at the drop of a question have long used controlled languages, which they often call taxonomies because they use them in hierarchical form. Lawyers have now found that the investment in developing a taxonomy, even for a single major piece of litigation, and in indexing the discovery documents, is much less than the continuing cost of paralegals (or attorneys) to do text searching over and over again -- sometimes for the same documents.

In addition to the economic objections raised by early database providers, was the objection that such organization might be too complex for the user -- but people have learned to use and to value other organized reference works: encyclopedias, back-of-the-book indexes, and even the yellow pages of a telephone directory.

Perhaps it is the idea of an intermediary language that is unsettling: "Why can't I go straight to what I want?" Yet nobody, on arriving in a strange city, would expect to go directly to the city hall (even knowing the address) without a map, or without a knowledgeable cab driver (the reference

librarian) who already knows the map.

Economic Considerations

The database preparation cost of a free- or full-text system is minimal, no more than what it costs to put the files into the computer. More terms are available for searching than are normally affordable in a controlled index language, and therefore, are likely to include more precise terms than can be included in the controlled index language. The searcher is thus offered a choice of great exhaustivity (range of topics) or great specificity (precise terms) depending on the manipulation of the search terms and the construction of the search statements and a search strategy.

Therein lies the main, and two-sided, difficulty. The searcher must trust the author to have included all of the terms necessary to describe the topic. There are databases in all fields in which a writer may not use the main topic term at all. When the *Washington Post* newspaper went online with its PostHaste system, it asked CDB Enterprises to construct a thesaurus so the articles could be offered in both indexed and full text forms. Their reason: a natural language search on the word MURDER might retrieve an article on July weather -- "It's murder out there"; and some articles never included a word for the main topic, because it was obvious. Indeed, in that summer, as we worked on their thesaurus, the golfer Arnold Palmer hit two holes-in-one on the same hole on two successive days in a major tournament, and the article describing this unprecedented feat never mentioned the word GOLF. Since writers are not consistent in their use of all the words that can be used to describe topics, the searcher must plan very carefully to include all synonyms and all necessary levels of specificity in searching. Further, simple Boolean intersection of natural language terms can produce imprecise results when two or more terms are present in a record, but without the semantic association expected by the user. (A search for window shades including the terms VENETIAN and BLIND might retrieve information on SIGHTLESS PERSONS IN VENICE.) Proximity searching can overcome this problem to some extent.

In other words, effort and time, and therefore cost, saved in database production is passed on to the searcher. Such a situation may satisfy a commercial database provider, by saving expense at the beginning of the process and increasing income at the point of use, but is not likely to satisfy the user. Full-text searching and controlled language searching are not the only access methods; they are used here only to emphasize two, possibly extreme, approaches. Further possibilities will be discussed at the end of the article.

It is not uncommon today for a database to contain records collected and indexed over a span of 50 years by a variety of index languages or, indeed, by none at all. Over time, the working language of a discipline changes and, eventually, reaches the point where a new, consistent index language is needed.

What is a Thesaurus?

A thesaurus, to a layman, is a fat book prepared by somebody called Peter Mark Roget and used by college students to enlarge their vocabulary when

writing term papers -- and, often and unfortunately, to vary the representation of the same concept from sentence to sentence. A thesaurus to an information scientist is a controlled set of the terms used to index information in a database, and therefore also to search for information in that database so the same concepts are represented by the same term. For many years in this country, thesauri were often presented as alphabetized lists of key terms, taken from the document to be indexed with references to and from other terms made as necessary. This traditional practice has changed in recent years to a more structured approach based on an analytical technique. Ironically, this means that the original misuse of the word "thesaurus" by information scientists, to describe purely alphabetical lists of terms (Roget organized his thesaurus by categories of knowledge, and included an alphabetized list of terms only as an index), has been amended so that it is now closer to a proper use of Roget's meaning to include both categorization and alphabetical listing.

"Thesaurus" is only one name for a controlled index language or one of its parts; "taxonomy", "classification", and "hierarchy" are others. "Keyword list" should really be used only for a list of terms taken directly from authors' original language, and "ontology" should not be used at all, because it means "the study of being" (Greek ONTO -- (being) + LOGIA (discourse) -- not a possible and mistaken back formation into ONTO (essence) + LOGOS (word) -- as the essence or existence of words).

There are three levels of thesauri:

1. Universal, like the Library of Congress List of Subject Headings, or the Library of Congress Classification scheme, or the Dewey Decimal Classification scheme;

2. Broad areas, like the Medical Subject Headings (MeSH) of the U.S. National Library of Medicine, or the Thesaurus of Engineering and Scientific Terms (TEST) originally from the Engineers Joint Council, or the Art and Architecture Thesaurus (AAT) supported by the Getty Trust; and

3. Specific areas, like the Transportation Research Thesaurus (TRT) administered by the Transportation Research Board of the National Research Council, or the ERIC Thesaurus on education.

The Power of Structured Language

This article makes an assumption about the nature and appearance of a thesaurus, even though many thesauri have been and will be created that use only some of these features, and in varied forms. The assumption is that a well-developed thesaurus is based on a recognition of a hierarchical structure, or a set of them, in which clusters of concepts that share a common characteristic are organized in families (called "facets"), and represented by natural language terms useful in the context in which the thesaurus will be used.

For example, Wood, Nylon, Steel, Copper, Wool all share the characteristic of being MATERIALS, whatever other characteristics some of them may share, like Combustibility. The terms constitute the beginnings of a

MATERIALS facet. Sometimes the terms in a facet can be divided into subfacets by secondary characteristics: Within the MATERIALS facet, Wood and Wool are ORGANIC MATERIALS; Steel and Copper are METALS; and Nylon is a PLASTIC. Facets and subfacets are then arranged as simple hierarchies of terms, from general to specific.

Facet procedure has many advantages. By organizing the terms into smaller, related groups, each group of terms can be examined more easily and efficiently for consistency, order, hierarchical relationships, relationships to other groups, and the acceptability of the language used in the terms. The faceted approach is also useful for its flexibility in dealing with the addition of new terms and new relationships. Because each facet can stand alone, changes can usually be made easily in a facet at any time without disturbing the rest of the thesaurus.

The faceted approach combined with the use of notational codes to represent the hierarchies is especially amenable to software applications. CDB Enterprises developed its own software for thesaurus construction: it ensures the integrity of the structure of hierarchies and references between them; it generates printed and on-screen displays of different thesaurus formats; and it even displays the thesaurus for point-and-shoot allocation of terms in online indexing. Nevertheless, the process of facet analysis and the construction and maintenance of a thesaurus is, and must always remain, essentially an intellectual endeavor.

A further and final benefit of the faceted approach becomes apparent in the use of the thesaurus: it is much easier for an indexer or searcher to understand a set of hierarchically organized facets as a conceptual map, which shows the precise level and set of associations of a term. Categories of related terms are easier to negotiate than a long list of alphabetized terms.

Typical thesaurus displays based on a faceted structure are:

- Explicitly hierarchical displays, sometimes called taxonomies,
- Alphabetical displays, showing references between related terms or references from unused synonym terms to the postable or preferred term,
- Rotated displays, in which each word in every natural language term is used as an index term to indicate the several phrases in which it appears.

Other Approaches

Enough has been said above to make it clear that full-text searching is fraught with the risk of overwhelming retrieval, and thus confusion, and that the development and application of controlled index languages (though extremely efficient in use) can be expensive enough to discourage authors or database providers. Are there alternatives?

Hypertext, with its links from an assigned word or phrase in a document to a related document is attractive, but it places on the author the same burden (perhaps heavier) as indexing the document using a controlled language. Increasingly, electronic information providers are using hypertext techniques to navigate their files, both within a file and across files, even

for ephemeral information like daily news extracts.

A transparent search thesaurus is a concept that has been discussed in information science research literature, but not yet implemented effectively. It would require the analysis of an extremely large collection of text in machine-readable form, possibly using an approach similar to Lauren Doyle's automatic indexing research in the 1960's on "semantic road maps" to generate a very large thesaurus with connections of varying degrees of intensity between terms. A search entered by a user would assemble all synonyms and tight semantic relationships in a string of OR relationships in the search statement, and join each string to other, conceptually separate strings in a group with AND relationships between each string. The combined search statement could then be used in full-text searching -- assuming that the text is susceptible to searching by the search engine and assuming that the search engine works at a speed tolerable to the user's expectations.

Conclusion

Perhaps the information environment and content, and the way people expect to approach them, prompt the access mode in a machine-sensible environment. Thus, scholarly and technical information, especially when organized as reports or reflective communication, is probably most susceptible to a combination of controlled index/search language and full-text searching, because the conceptual structure and the vocabulary are likely to be well organized. Indeed, the authors of such information are motivated to organize and even index their work because of their intention to disseminate the information. Digested information like news bulletins, current awareness commentaries, and even sales catalogs, directories and handbooks are better suited to hypertext techniques. Undigested and unorganized information that is often the author's self-serving imposition on the Web probably deserves only a full-text search engine general enough to handle more or less anything. Like the old computer produced rotated term indexes that relied only on the words in the titles of printed documents, the searcher has to work harder and longer than the author or the system, and may find only a little of what the authors produced and did not identify carefully enough.

Information is as accessible, for retrieval and for comprehension, as its author or its publisher makes it. There is a burden of effort in information storage and retrieval that may be shifted from shoulder to shoulder, from author, to indexer, to index language designer, to searcher, to user. It may even be shared in different proportions. But it will not go away. Once we realize and face this, we may begin to make sense of the wealth (or waste) of electronic information.

References

David Batty, "Words, words, words" in "Database Design" column, Database, December 1988, pp. 109-113.

David Batty, "Theasurus construction and maintenance: a survival kit." Database, February 1989, pp. 13-20.

David C. Blair and M. E. Maron, Communications of the ACM, March 1985, 28:3, pp. 289-299.

David C. Blair and M.E. Maron, Information Processing and Management, 1990, 26:3, pp. 437- 447.

Gerard Salton, Communications of the ACM, July 1986, 29:7, pp. 648-656.

David Batty is President of CDB Enterprises, Inc., a consulting company that specializes in library and information system design, information storage and retrieval systems, and especially the controlled index languages that drive them.

Copyright © 1998 David Batty

Top | Magazine
Search | Author Index | Title Index | Monthly Issues
Previous Story | Next Story
Comments | E-mail the Editor

D-Lib Magazine Access Terms and Conditions

hdl:cnri.dlib/november98-batty

Faceted Classification and Logical Division in Information Retrieval

JACK MILLS

ABSTRACT

THE MAIN OBJECT OF THE PAPER is to demonstrate in detail the role of classification in information retrieval (IR) and the design of classificatory structures by the application of logical division to all forms of the content of records, subject and imaginative. The natural product of such division is a faceted classification. The latter is seen not as a particular kind of library classification but the only viable form enabling the locating and relating of information to be optimally predictable. A detailed exposition of the practical steps in facet analysis is given, drawing on the experience of the new Bliss Classification (BC2). The continued existence of the library as a highly organized information store is assumed. But, it is argued, it must acknowledge the relevance of the revolution in library classification that has taken place. It considers also how alphabetically arranged subject indexes may utilize controlled use of categorical (generically inclusive) and syntactic relations to produce similarly predictable locating and relating systems for IR.

1. INTRODUCTION

As a memorable aphorism prefacing his novel *Howard's End,* E. M. Forster gave simply "Only connect." It could claim to be the finest, even though briefest, definition of intelligence we have. To understand anything, whether it is the operation of a complicated mechanism or the complex social factors that underlie almost any human situation, understanding it means seeing the connections. The basic intellectual instrument we use to do this is classification. It is appropriate that libraries, which seek to organize ev-

Jack Mills, Editor, Bliss Bibliographic Classification (BC2) c/o Bliss Classification Association, The Library, Sidney Sussex College, Cambridge, CB2 3HU, United Kingdom
LIBRARY TRENDS, Vol. 52, No. 3, Winter 2004, pp. 541–570
© 2004 The Board of Trustees, University of Illinois

erything in the way of recorded human knowledge should find explicit classification as central to their organization.

1.1. Indexing and searching

Indexing and searching are the two fundamental operations in retrieval. The usual situation in the library is that the librarian prepares the scene for retrieval by indexing each document (assigning to them retrieval handles such as classmarks, subject headings, etc.). Searching may then be done directly, by examining the documents on the shelf or vicariously via their surrogates in the catalog. Although the term "indexing" is used with various connotations, especially ones involving terms in alphabetical order, the central meaning of pointing out or indicating describes exactly what librarians do when, in response to any enquiry, they indicate where the inquirer may best begin looking and, perhaps, where they might next look should the first search prove inadequate. This function is neatly summarized in cataloging theory as one of locating and relating.

1.2. Classification

This is the most fundamental operation in indexing. In its broadest sense, it is the action of recognizing and establishing groups of classes of objects, the subclasses and members of which all manifest (even though in different ways) a particular characteristic or set of characteristics. The different kinds of shared characteristic(s) used to define a class for retrieval have been called index devices (Cleverdon et al., 1966). Library classification, via shelf order and the classified catalog, uses a number of different devices; two of these reflect the sort of class definition usually understood by the term "classification"—those defined by generic and whole-part relations; but coordination (combination), synonym control, role indication (by inclusion of terms in facets defining their relation, such as agent, property), and some confounding of word forms (via their adjacency in the A/Z index) are also prominent. Mechanized retrieval systems developed a number of less direct devices, e.g., an extended confounding of word forms and oblique ways of defining a set of documents sharing the same subject content such as is found in citation indexing. Electronic systems have now extended these oblique forms of class definition (see Section 3.5).

2. What Is Classified in the Library

Library materials physically are the object of relatively rudimentary classification in that significantly different physical forms are separately housed and may be separately indexed. However, in nearly all cases it is their content which is their ultimate justification and the problems of information retrieval (IR) are paramount. Whether this content is best described as information or knowledge is best left to the philosophers. Early writers on library classification tended to use the term "knowledge" as the object of classification and retrieval. A dissident voice at the beginning of the last

century, when Bliss began opting firmly for a knowledge basis, was Wyndham Hulme (1911–1912). Hulme distinguished mechanical classification from philosophical and claimed that library classification belonged to the first kind. He coined a term "literary warrant" and described library classification as the plotting of areas preexisting in literature. This was, in fact, not a bad description of what the Library of Congress was doing in many of its classes but interpreting preexistence as being what they held in stock. When we consider content only, a major distinction is found in all general libraries (and some special) between subject content and what we may call, for lack of better words, "imaginative content." Much discussion of the exact nature of information reflects the unease over the use of the term "information retrieval" when it is clear that the content of a significant class of documents is not defined sensibly as information. The term "knowledge" appears to be somewhat more receptive to the inclusion of imaginative works than the term "information."

2.1. Subject content and imaginative content

The latter has received attention by and large only in respect of fiction (Beghtol, 1994; Hjørland & Albrechtsen, 1999). But the well-established dichotomy between fiction and nonfiction is somewhat misleading. Fiction is only one example of imaginative content; the latter includes also other literary forms (poetry, drama), all musical compositions, and all forms of the visual arts that can form the content of a record (e.g., a folio of paintings). If fiction offers viable characteristics of division whereby it can be organized, the same characteristics, in principle, should be applicable to all of them. The folio of paintings (say) might be classified by creator or by place (French paintings, etc.) or by period (twentieth century, etc.). But the above characteristics represent logical categories that are common to all kinds of record content. Music scores are classified by instrument (vocal, instrumental, etc.) and only secondarily by creator. But some characteristics might be thought to be special to imaginative works. For example, the new Bliss Classification (BC2) (see Section 5.2) includes in its Properties facet of Class W The Arts such terms as "didacticism, parody, sentiment, realism" and in its Elements facet terms like "symmetry, rhythm, symbolism, fantasy." By the process of specification (see Section 7.3), this allows imaginative works to be classified as didactic, parodic, sentimental, symmetrical, rhythmic, symbolic, fantastic, and so on. But many of these could also characterize subject content (in individual behavior, social behavior, technological work, etc.). In practice, much of the classification of imaginative works, especially fiction, is by subject content. But iconographic art (and its opposed nonfigurative or abstract art) inevitably uses the concepts making up a subject classification itself. The Subjects of art facet in BC2, for example, makes direct use of the whole classification, which gives a comprehensive and predictable order. Insofar as the classification of imaginative works

raises problems of cross-disciplinary and cross-cultural influences, it does not differ essentially from subject classification. The rules developed for the systematic handling of such relationships (see Sections 5/7) are as applicable to imaginative content as they are to subjects. Where imaginative content does present a special problem is that the categorization of a given imaginative work by some or many of the characteristics available would most likely be very subjective, and this factor almost certainly limits the degree to which they are practically feasible. But this does not mean that the rules present a rationalistic bias when applied to imaginative works, only that they are essential to the aim of achieving predictability in location, whatever the content of the record.

2.2. Common-sense view

The interpretation assumed in this paper of what exactly is classified may be described, for better or for worse, as the common-sense view of most librarians. The object of attention in library classification is the content of records; they will have embedded in them, to varying degrees, matters of fact (as Hume would say, in a famous phrase that, incidentally, begins with "When we run over libraries . . .") accompanied by considerations of analysis, discussion, prediction, opinion, and other matter (much of which might be considered to fall within the category of "relations of ideas") and other less concrete matter that may or may not be deemed worthy of inclusion in the index description. But if it does appear, it will be susceptible to logical division.

3. Levels of Indexing in the Library

A reader so far may have assumed that the catalog is the form par excellence of an index to the library collection and the prototype of indexes to larger collections and networks. This is not quite true. A library is indexed for retrieval at three levels: the systematic order of documents on the shelves (assuming complete or partial open access), the A/Z index to the classification governing the systematic order, and the catalog.

3.1. Shelf order

This is scarcely ever mentioned in the literature on retrieval, being treated very much as a poor relation, if not a terminally ill one. This is most unfortunate, since it is the very first index to the resources of the library for the great majority of library users and in many cases the main or even only one. Although this level of retrieval may be regarded as small beer and not deserving much attention, the special demands it makes because of its limitation to a single, linear order has had an important effect on the development of the theory of library classification. The limitation to a linear sequence throws into sharp relief a crucial property sought in indexing—that of predictability as to the location of any given class of information. The physical document can only go in one place. But the concepts that

define the class represented by that one place are in most cases multiple, e.g., a class represented by the rubric Bone—Cancer—Therapy—Radiography could legitimately go in any of twenty-four different places, everyone of them making sense. The expectations of users reflect this. A radiographer would like to see it under medical radiography; the cancer specialist would like to see it under cancer, and so on. The implication is clear. The classification must have comprehensive rules governing the order in which the different component parts of a compound subject are to be taken when locating a class. This does not depend in any way on the specificity of the index descriptions given to the documents; even if the classmark locating it is not specific (i.e., reflects "broad classification") the librarian and library user still need to know where it will go—under skeletal system, or therapeutics, or radiography, or cancer.

3.2. The A/Z index to the classification

The relative index that Dewey provided for his classification has been an outstanding example of this indexing component since the scheme was first published in 1876. It intuitively recognized that the central weakness of the classified index represented by the shelf order is that it distributes many subject concepts over many fields according to the rules for combination already mentioned. So, for example, literature on children will be scattered as a result of its subordination to different containing classes—medicine, psychology, education, welfare, and so on. Hence, Dewey's (1985) term "Relative Index" and the general use of the term "distributed relatives" to describe the situation.

3.3. The catalog

This consists of surrogates representing the records themselves, each surrogate containing, to a greater or lesser degree, a bibliographical description and rubrics to act as retrieval handles (indications of its subject, author, title, etc.). It has two central functions: first, as an inventory of the library's holdings; second, it provides for multiple access in searching (by author, title, form, or subject). Accessing by subject presents the central problem, and it is the subject catalog that is considered below.

3.4. Precoordinate indexes

Apart from a few special collections, this was the only form of subject catalog used until the 1950s. The term refers to the handling of compound subjects, which constitute the vast majority in the literature. The constituent terms that in combination (coordination) describe the subject are coordinated in the subject heading or classmark in anticipation of the needs of searchers. Compounding immediately raises the problem of distributed relatives; this problem, absolutely central to shelf order, continues to be central to the organization of the surrogates also, despite their much greater facilities for providing multiple access. How the separate concepts needed

to describe the compound subject are linked depends on the relationships subsisting between them, and these, in turn, determine the search strategies for locating the information sought. The problem of distributed relatives that this poses can be ameliorated (but never completely resolved) by making multiple entries for a document with a compound subject so that a separate entry appears directly under each of its major constituent concepts. For example, the document referred to earlier might get a separate entry under each of the four constituent concepts: Skeletal system, Cancer, Therapy, and Radiography (but omitting separate entries for the other twenty permutations theoretically possible). Such permutation is standard practice in libraries using the Universal Decimal Classification (UDC), whose notation particularly provides for it. Such permutation of multiple entries is rarely found in the alphabetical subject catalog. Notably, most subject cataloging takes as its unit a complete and discrete record (a book or article), and its classification and indexing involve a process of summarization. The subject description is of the record as a whole, and this determines its position.

3.5. Postcoordinate indexes

The development of mechanical aids to indexing (e.g., peek-a-boo, machine punched-cards) from the 1920s onward saw the removal of the need to summarize the overall content in a single precoordinated subject description. Now, only single constituent terms were assigned, and their combination to form a search request for the subject concerned was left to the search stage. This system was called postcoordinate indexing since the coordination appeared after the indexing step, requiring less effort since it moved the burden onto the searcher. The absence of recognized relationships could result in ambiguity, e.g., a search for fertilizers for sugar beet by the simple coordination of Sugar and Beet and Fertilizers would also produce documents on the use of sugar-beet tops as fertilizers. This led to the reintroduction of classification at the indexing stage in the form of role indicators and other devices that are implicit in the precoordinate index.

Mechanical aids were soon supplanted by electronic systems, and a still more drastic change in indexing practice followed. With the development of networks for electronic retrieval, the economic burden presented by the prior indexing of individual records (typically, for services operated commercially) became prohibitive. Now, it was not just a case of abandoning the intellectual precoordination of index terms but the abandoning of preindexing altogether. Reliance was to be entirely on keywords found in the record and recognized by electronic searching. Indexing devices developed by librarians can only be used indirectly, by assisting the framing of requests to search engines. The limited discriminatory powers of keywords, with all their attendant ambiguities in the unruly natural language, were now supplemented by new index devices, with machines operating on the

relatively raw text of the documents. All of them are based on the measurement of relatively artificial characteristics of documentary texts, such as frequency of occurrence of particular words, contiguity of particular words, etc., using statistical techniques and mathematical algorithms. These are deemed sufficiently correlative to conceptual meanings to form classes allowing searches defined conceptually. They constitute new index devices, but they are still classificatory in operation, establishing subclasses of the total store identified by the parameters of the technique used. They are not assigned by an indexer but must utilize the computer programs of the store's service provider.

The shift from IR from stores of limited size, in which trained librarians have prepared the field for searching by the prior indexing of materials, to much larger stores in which there has been only minimal preparation of the field has important implications for the relationship of libraries to information science. The cognitive processes connecting the producers of texts stored and the would-be recipients of the knowledge stored in the texts are the subject of much current research. However, the highly structured maps of knowledge developed by modern faceted classification apparently have considerable potential in assisting these processes.

4. Indexing in the Library Today

The inroads on the librarian's time made by the need to master rapidly developing computer techniques has had a particularly unfortunate effect on the curriculum of library schools, where the study of the organization of knowledge has been eroded just when the need for it has become greater. The information explosion led, inter alia, to the development by librarians of greatly improved index languages, largely based on facet analysis. The relevance of these to the future of the profession assumes two things: first, that the library will continue to be an integral part of our culture and that reports of the birth of the paperless society have been greatly exaggerated; and second, and following from the above, we have an obligation to seek the best possible ways of facilitating its work. The development of logically structured classifications covering the whole of knowledge is still unique in the field of LIS. These provide detailed maps of knowledge to assist in the searching of stores of records and can be used as the basis of, or valuable supplements to, numerous other retrieval languages.

5. The Design of a Modern Library Classification

Two conceptual areas must be distinguished: general classifications covering all knowledge and special classifications restricted to a specific field. The significant developments in classification design claimed above refer primarily to the second area and will be considered in detail under that. Here, some distinctive features of a general classification are considered.

5.1. General classifications

Remember that all special classifications need to draw on a more general one, often extensively. Another reason why IR cannot afford to ignore the concept of a general classification is that it alone can provide a bird's-eye view of the whole field of knowledge, offering a comprehensive context within which searches in a very large store can be framed. How the main classes (a loosely defined but reasonably well-understood concept) are handled within a general classification is the main theme of this paper. But whereas the central feature of the faceted special classification is its rigorous observance of the rules of logical division (see Sections 5.3/7), this cannot be said to apply initially to a general classification. If the first step in establishing what are loosely called its main classes were to be the division of the whole field of knowledge by applying explicit characteristics of division, the only feasible contenders would be of the nature of fundamental categories. The earliest and best-known set of such categories is seen in those advanced by Aristotle. Some of these are ostensibly feasible as constituting the initial divisions of the whole field of knowledge, e.g., substance, quantity, quality, place, time, and action. Such a first step has not been attempted by any of the general library classifications produced since Dewey's annus mirabilis in 1876, although something like it was attempted by the Subject Classification of the British librarian James Duff Brown (Brown, 1939/1906) with its quadruple division into Matter, Life, Mind, Record. Brown's scheme was notorious in its day for its subordination of music to sonics in physics—an example of its attempt to ignore disciplines as a primary level of division. What did emerge, with a relative unanimity that is not really surprising, was an initial division into main classes reflecting the division of labor—intellectual, imaginative, and practical. The division of labor is a fundamental feature of society, which is itself the producer of the knowledge in the records that are the objects of IR. It is manifested in every sphere of society, including academia as well as in the practical production of material wealth. The term "discipline" is frequently used to refer to these specialized fields, but is ambiguous insofar as a truly main class (e.g., the natural sciences) is usually susceptible to logical division into subclasses that are themselves known as disciplines.

The particular notion of the fundamental forms of knowledge that underpin main classes has received significant attention by Langridge (1976), who has drawn extensively on the work of a number of philosophers, particularly that of Hirst (1974) and of Phenix (1964) in the philosophy of education. Of particular significance is the distinction Langridge draws between the forms of knowledge on the one hand and the objects of knowledge (the phenomena they examine) on the other. The order in which main classes might appear became a particular focus of attention in the work of Bliss (1929, 1933), and a modified form of the general order he advocated is considered in Section 5.2.

A common criticism of the viability of any schema of universal knowledge is that the interaction of existing fields tends to dissolve their boundaries. While this interaction and its tendency are indisputable it does not invalidate the search for relatively permanent structures. Work on BC2 (Mills & Broughton, 1977–) has not found the great waves of new specializations an insurmountable obstacle. With enduring principles like gradation and integrative levels, together with highly practical principles such as the subordination of means to ends to reflect the concept of purpose or end-product to determine citation order within a given class (see Section 8.2), the predictability in the location of quite intricately mixed specializations is ensured. For example, modern forensic science draws on chemical analysis, molecular biology, and any number of medical specializations, but the purpose it serves—to validate the evidence in legal processes—determines its location in the law class with high predictability.

5.2. Two modern general classifications

The Colon Classification (Ranganathan, 1960) is not included here; its significance is primarily that it pioneered faceted classification and provided an experimental test-bed for its development. But its main-class order is quite conventional and offers no solutions to the problem of general classifications per se. The Broad System of Ordering (1978), or BSO as it is usually called, was first designed as a switching language—i.e., an intermediary through which other classifications could translate into each other. Its lack of detail stems from the fact that it was initially based on an institutional warrant—i.e., of subjects displaying institutional organizations underpinning them rather than on the much larger literary warrant of library collections. One feature is the break it makes with the generally recognized fields of knowledge, e.g., it has separate general classes for important concepts normally distributed under different contexts, e.g., Communication and information, Management, Human needs. It also has a Phenomena class (see Section 5.1) for works that cannot be accommodated in any of the largely disciplinary main classes, which are in BSO all fully faceted. It has also been very influential in the development of the next system, BC2.

For historical reasons, as well as theoretical ones, the BC2 (Mills & Broughton, 1977–) has largely taken the main-class order of the original Bibliographic Classification (Bliss, 1940/1953). This order reflects the Comptean principle of gradation and that of integrative levels (Feibleman, 1954; Foskett, 1961). The major sequence these give is modified in a few respects, as is shown in the outline in Appendix 1. BC2 has completely restructured all the individual classes, and each class is now fully analytico-synthetic in structure and notation. It is now virtually a new general classification and constitutes the most detailed, fully faceted general classification in existence. For this reason it is used in this paper as an exemplar of faceted structures, which are now (from the work done on it) seen to be appli-

cable to every field of knowledge. Like BSO, it also includes a separate Phenomena class, in which the order of phenomena closely follows the main-class order and uses the principle of unique definition to determine the location of multidisciplinary works on a given phenomenon. An outline of the system is given in Appendix 1.

5.3. Faceted classification of a subject field

This has been the major development in classification for IR in libraries in the past fifty years, although its first formulation was in the work of Ranganathan. Although, curiously enough, Ranganathan never referred explicitly to the fact, the fundamental feature of his Colon Classification is that it divides any given subject in accordance with the rules of logical division. But logical division is not the whole story. The work on BC2, covering every field of knowledge, clearly has shown that the design of a special classification requires recognition of six fundamental steps. These steps must of necessity be taken in the same order, since each step depends on the completion of the previous one. Only the first two use logical division; the other four use extralogical procedures. The steps are easily summarized:

5.4. The six fundamental steps in design are

- Division of the subject into broad facets (categories);
- Division of each facet into specific subfacets (usually called arrays, following Ranganathan);
- Deciding the citation order between facets and between arrays;
- Deciding the filing order between facets and between arrays and the order of classes within each array;
- Adding a notation;
- Adding an A/Z index.

5.5. The role of logical division

Before considering each of these steps in detail, the general role of logical division, which governs the crucial first two steps, must be noted. The rules of logical division, developed more than two millennia ago, are admirably brief:

- Only one characteristic of division should be applied at a time;
- Division should not make a leap; steps should be proximate;
- Division should be exhaustive.

The first and crucial rule is purely one of conceptual analysis and doesn't depend on practical considerations. The second and third rules involve to some extent subjective practical considerations as to the size of vocabulary to be accommodated and the degree of specificity with which compound classes are to be described. They are manifested only at the level of arrays (see Section 7). Observance of the first rule is the hallmark of faceted classification; a classification that fails to observe it rigorously throughout

The International Information & Library Review (2004) 36, 47–54

ELSEVIER

The International Information & Library Review

www.elsevier.com/locate/iilr

The development and structure of the Chinese Thesaurus for subject indexing

Wenxian Zhang*

Archives and Special Collections, Olin Library, Rollins College, Box 2768, Winter Park, FL 32789, USA

Received 26 August 2002; received in revised form 12 August 2003; accepted 13 October 2003

KEYWORDS
Thesauri;
Chinese Thesaurus;
Chinese information storage and retrieval;
Chinese Classified Thesaurus;
Subject indexing

Summary In the information age, a thesaurus still remains as an important retrieval tool. The purposes of this article are to review the development and nature of the *Chinese Thesaurus* (*CT*), and to examine its organizational structure and functions in Chinese information indexing. As the largest, and most comprehensive indexing retrieval devices for Chinese language, the *Chinese Thesaurus* and later the *Chinese Classified Thesaurus* (*CCT*) have played key roles in the standardization of Chinese retrieval language, and contributed greatly to the modern development of knowledge organization and information processing in China.
© 2004 Elsevier Ltd. All rights reserved.

Introduction

In an information-based society, effective knowledge organization and retrieval are important parts of our lives. Thesauri, along with classification systems and subject headings, are essential tools for organizing and retrieving complex information through the use of controlled languages or predetermined vocabularies. For the purpose of this article, a thesaurus is defined as a vocabulary of controlled indexing language, formally organized so that certain relationships between concepts are made explicit, to be used in information retrieval systems, ranging from the card catalog to the Internet. The primary purpose of a thesaurus is for information retrieval, and secondary purposes include aiding in the general understanding of a subject area, providing "semantic maps" by showing inter-relations of concepts, and helping to provide definitions of terms (Aitchison, Gilchrist, & Bawden, 2000).

Given the usefulness of a thesaurus in information organization and retrieval, this paper attempts to review the development and nature of the *Chinese Thesaurus* (*CT*), and to examine its organizational structure and functions in Chinese information indexing. First published in 1980, the *Chinese Thesaurus* (Han Yu Zhu Ti Ci Biao, 1980) is the largest, and most comprehensive indexing retrieval device for Chinese language. As a result of collective efforts by thousands of Chinese library and information professionals, its development reflects the unique social and cultural backgrounds of China today. Discrepancy exists in the compilation of the *CT*, its relatively low specificity of subject terms also makes it impractical for a wide adoption among special libraries and information institutions in the country. Despite its drawbacks, *CT* and later the *Chinese Classified Thesaurus* (*CCT*), along with the *Chinese Library Classification* (*CLC*), have played a key role in the standardization of Chinese retrieval language, and provided a solid foundation for compiling subsequent thesauri in China.

*Tel.: +1-407-646-2231; fax: +1-407-646-2122.
E-mail address: wzhang@rollins.edu (W. Zhang).

1057-2317/$ - see front matter © 2004 Elsevier Ltd. All rights reserved.
doi:10.1016/j.iilr.2003.10.001

The development and the nature of the Chinese Thesaurus

Before the 1960s, very few individuals or institutions in China compiled subject catalogs and thesauri. The *Hong Kong Subject Headings of Science and Technology* (Ke Ji Zi Liao Zhu Ti Biao) in 1964 was the first one taken into practice (Zhang et al., 1996). In light of emerging computer technologies and rapid surge of scientific information, the Chinese government initiated Project 748 in 1974, a systematic project on Chinese information processing. As a part of the project, the *CT* was compiled by the Institute of Scientific and Technical Information of China (Zhong Guo Ke Xue Ji Shu Qing Bao Suo) and the National Library of China (Bei Jing Tu Shu Guan). More than 1370 information professionals of various subject fields from over 500 institutions participated in this collective effort (Lu, 1993). Published in 1980 and with 91,158 preferred terms and 17,410 non-descriptors, it has become the largest, and most comprehensive indexing retrieval device for Chinese language. Comparing with other major thesauri available in the late 20th century, it contains a substantial number of terms and deals with a wide range of disciplines and technologies. The *CT*, along with the publication of the *CLC*, marked a new era in the Chinese librarianship of the 20th century. In 1985, the *CT* won the second-class prize of the National Award for the Advancement of Science and Technology (Zhang et al., 1996, p. 33).

The goal of the *CT* is to provide for Chinese information processing, a working vocabulary in all fields of sciences and technologies. All of the 108,568 terms are in Chinese. They are included in three massive parts and 10 volumes: social sciences and humanities (Part I, Volumes 1–2); natural sciences and technologies (Part II, Volumes 3–9); and appendixes (Part III, Volume 10). Each of the first two parts has a thesaurus of alphabetical terms along with hierarchical index, subject category index and English–Chinese bilingual index.[1] Appendixes include countries and regions of the world, geographical areas, and lists of organizations, agencies and important persons. In addition, there is also an abridged edition with 12,995 terms selected from the *CT*, covering both social and natural sciences and containing only an alphabetical list of descriptors.

The *CT* was designed with multiple functions. As a comprehensive tool for Chinese indexing and processing, it was compiled for use in not only library catalogs but also published alphabetical indexes as well as computerized retrieval systems. To this end the Chinese library and information professionals strived to develop a general, yet practical thesaurus which would cover broad subject fields for indexing Chinese publications. They hope that the successful implementation of the *CT*, along with a number of specialized subject thesauri, would ultimately lead to the formation of a single, uniform thesaurus for information processing in China. Through this effort, the country's information retrieval language would be standardized and eventually, a Chinese national subject retrieval system constructed.

Structural arrangement of the *CT*

The main part of the *CT* is the alphabetical list of descriptors and other related terms. As defined by ISO 2788,[2] an indexing term is the representation of a concept. A descriptor, also known as preferred term, is a term used consistently when indexing to represent a given concept (Aitchison et al., 2000, p. 17). In the *CT*, the descriptors for social sciences and humanities are largely selected from vocabularies of philosophy, politics, economics, culture and other subject areas. They include technical terms, special concepts, school of thoughts, political viewpoints, historical events, conference documents and other specialized nouns. Lexis of natural sciences and technologies includes terms regarding various subjects, materials, methods, techniques and functions. In order to meet the needs of both manual and computerized indexing, many compound terms are listed, either pre- or post-coordinated.

The order of listing in the *CT* is based strictly on the alphabetized Chinese pronunciation of a descriptor, regardless of the number of strokes and radicals of Chinese characters.[3] By 1991 in the

[1] It should be noted that alphabetical list here refers to a system of standard Chinese pronunciation based on the International Phonetic Alphabets. As Chinese is an ideographical language in which characters, not alphabets, are the basic building blocks, in the 1950s the Chinese central government devised the Pinyin system to assist children and foreigners to pronounce the language correctly. Since then, Pinyin has been widely used for standardized translation of specific Chinese names and places, as well as listing and indexing words and phrases of Chinese language.

[2] ISO 2788 refers to the international standards for the Establishment and Development of Monolingual Thesauri issued by the International Organization for Standardization in 1974. It was later revised in 1986 as ISO 2788:1986.

[3] As radicals are the basic components in most Chinese characters, radicals combined with total number of strokes in a given character used to be the most commonly utilized system of listing and indexing Chinese Language before the invention of Pinyin.

revision of the science and technology segment of the *CT*, the alphabetized listing arrangement was further modified to accommodate the orders of tone designations and radical strokes so that terms with the same Chinese characters would be grouped together.[4] The actual listing includes a descriptor, its Chinese pronunciation and English translation, category number, scope note if any and relationship with other terms. For instance, the term of "Information Retrieval" is listed as:

Qing Bao Jian Suo	07k
Information Retrieval	
D	Document Retrieval
F	Subject Retrieval
	Retrospective Retrieval

In constructing the *CT*, noun forms or gerunds are used wherever possible. Direct entry is preferred to inverted entry where keywords are moved to the beginning of preferred terms. Synonyms, slight variations of same or similar words or phrases, abbreviations, quasi-synonyms and near-synonyms are regarded as equivalent terms of preferred descriptors. This relationship of equivalence structure is basically displayed in the following ways:

Y (Yong)—Use: indicates a main term or a descriptor that is to be used in indexing and searching;

D (Dai)—Used For: refers to a descriptor to be used from a term or non-descriptor that may not be used in indexing;

C (Can)—Related term: for additional references that are closely related to the main concept but less rigorously than the narrower or broader term.

Similar to an English thesaurus, additional methods are also employed in the *CT* to express hierarchical relationships:

F (Fen)—Narrower Term: always included within the main descriptor, refers to a term one level down the hierarchy;

S (Su)—Broader Term: always includes the main term concept, refers to a term one level up the hierarchy;

Z (Zu)—Top Term: main concept at the top of the hierarchy.

In addition to the main thesaurus of alphabetical terms, subject category index, hierarchical index and English–Chinese bilingual index are included in the *CT* to provide multiple ways of indexing and retrieval. The subject category index lists all descriptors and non-descriptors according to their subject disciplines. Built on the two main categories of social and natural sciences, the whole index is divided into 58 major classes, 675 classes and 1081 sub-classes. On average, each major class is further divided into 12 classes. The number of terms listed under each major class depends largely on its subject areas. For instance, there is a total of 4113 terms listed under 05: Economy. The listing is from general to specific, similar to a library classification schedule, and multiple entries are allowed. Each major class has a two-digit designation ranging from 01 to 92. There under, a mixed notation of numbers and letters is adopted following enumerative principles. The detailed and in-depth subject category index makes the *CT* stand out among major thesauri in the world. Table 1 below displays the 58 major classes of the *CT*.

The hierarchical index displays the structural relationship of super-ordination and sub-ordination among descriptors. In a logically progressive sequence, the super-ordinate term represents a class or whole, and the subordinate terms refer to its members or parts. There are more than 3700 families of descriptors in the *CT*, which comprise about 80% of all the terms in the vocabulary. The smallest families contain only two terms while the largest ones have more than 2000 terms. Consequently, some families have up to nine levels of hierarchy.

The third indexing tool in the *CT* is the English–Chinese bilingual index. Both preferred and non-preferred terms are translated, some descriptors are listed with multiple translations. Since the English terms are also provided under the main list, the *CT* has certain characteristics of a bilingual thesaurus. The index serves as a dictionary of standardized translations for both English and Chinese terms. It assists users in identifying needed English subject headings as well as directly searching original English publications.

The last volume of the *CT* consists of four tables: countries and regions of the world (1100 terms), geographical areas (361 terms), lists of organizations and agencies (1900 terms) and important persons (4765 terms). The first table includes all countries in the world, their major cities and special administrative regions. Names of Chinese districts include provinces, autonomous regions and major cities. County names are excluded. A few distinctive towns and villages are listed in the main body of the thesaurus under various disciplines such as "Ban Po Village" in archeology. Ancient place names, monuments and famous historical sites are also included in the main thesaurus. The second table contains the world's geographical names,

[4]Based on the Pinyin system, the actual pronunciation of Chinese characters is determined by a combination of phonetic alphabets with four different tone designations—flat, up, down-up, and down.

Table 1 List of major classes in the subject category index.

01	Marxism, Leninism & Mao Zedong Thoughts
02	Philosophy
03	Political Science
04	International Relations
05	Economy
06	Military Affairs
07	Cultural Affairs
08	Education
09	Sports
10	Linguistics & Language Scripts
11	Literature & Arts
12	History
13	Ethnics
15	Psychology
20	Social Sciences, General Concepts
30	Mathematics
31	Mechanics
32	Physics, Crystallography
34	Chemistry
35	Astronomy
36	Physical Geography
37	Geology
39	Surveying
41	Geophysics
43	Meteorology
44	Oceanography
45	Biology
47	Medicine
49	Agriculture & Forestry
51	Engineering Physics
52	Nuclear Technology
53	Energy & Power Engineering
54	Electrical Engineering
56	Electronics
57	Telecommunication Engineering
58	Automation, Computation & Computers
60	Navigation Technology
61	Light Industry
62	Chemical Industry
63	Petroleum & Natural Gas Industry
64	Mining
65	Metallurgical Industry
66	Metals, Metal Processing & Equipment
67	Mechanical Engineering
68	Civil Engineering & Urban Construction
69	Hydraulic Engineering
70	Construction Material Engineering
71	Transportation Engineering
72	Vehicle Engineering
73	Vessel Engineering
74	Aviation & Space Technology
75	Weaponry, Military Equipment & Technology
81	Metrology & Instruments
82	Experiments & Testing Equipment
83	Common Technologies & Crafts
87	Material Science
91	Environmental Science
92	Natural Sciences, General Concepts

including: mountains, rivers, lakes, oceans, islands, plains and basins. Terms such as Himalaya, Sahara Desert, Arctic Ocean, Korean Peninsula, Black Sea, Yellow River and Dzungarian Basin are included. In the third table, the organizations and agencies are usually listed under their full names. However when an institution such as NASA is better known for its abbreviation, the acronym is listed as a descriptor and cross-references are established. Finally, the significant individuals in world history are listed under the standardized Chinese translation of their last names, following with first and middle name initials and life spans in parentheses, e.g., Copernicus, N. (1473–1543). In sum, the four tables in appendices provide users not only additional terms, but also useful reference tools. They are important components of the *CT*.

Compatibility and revision of the *CT*

The inclusion of descriptors in its library catalog card services by the National Library of China has substantially raised the national awareness of the *CT*. However in practice, *CT* has only been fully explored by some large, comprehensive libraries in China for a limited period of time. Because of its long revision cycle, in some occasions it is regarded merely as a reference vocabulary instead of an indexing tool. Its relatively low specificity of subject terms also proved impractical for use by many special libraries and information institutions in the country. Nevertheless, the *CT* has provided a solid foundation for compiling subsequent thesauri in China. Shortly after its publication, it has been officially recommended that all abstract services should take the *CT* as their model for indexing (Zeng, 1986). Based on this principle, many institutions that participated in the development of the *CT* soon began to compile specialized thesauri covering almost all the disciplines. Hence in the field of information organization and retrieval, China has experienced a period of 20 years of rapid advance. Of more than 100 subject thesauri compiled since 1980, several of them are listed below:

Chinese Thesaurus of Chemical Engineering (1983): by the Institute of the Scientific & Technical Information, Ministry of Chemical Industry with 19,677 terms;

Chinese Thesaurus of Geology (1984): by the Library of Geology with 7063 terms;

Chinese Thesaurus of Petroleum Industry (1985): by the Institute of the Scientific & Technical

Information, Ministry of Petroleum Industry with 7588 terms;

Chinese Thesaurus of Urban & Rural Construction (1987): by the Development Center for Architectural Technology of China with 10,267 terms;

Chinese Thesaurus of Rolling-Stock (1987): Si Fang Rolling-Stock Institute, Ministry of Railway with 3158 terms;

Chinese Thesaurus of Railway (1988): by the Institute for Railway Information with 4529 terms (Lu, 1993).

The issue of standardization and compatibility of various thesauri has also been addressed in China. Despite a late start in thesauri development compared to other major Western countries, China benefits greatly from experience abroad in the standardization of indexing languages in the world (Zeng, 1990). Corresponding to ISO TC46, the National Technical Committee for Standardization of Information Documentation set up its 5th subcommittee in 1979, which actively engaged in standardization of information organization and indexing languages in China. As a result of their work, the *CT* was adopted as a national standard in 1980, same year of its publication (Hu, 1994). Other acts by the subcommittee in the formulation of national standards included (Zhang et al., 1996, p. 34):

Guidelines for Subject Construction and Term Indexing (Wen Xian Zhu Ti Biao Yin Gui Ze, GB3860-1989);

Guidelines for the Establishment and Development of Chinese Thesauri (Han Yu Xu Ci Biao Bian Zhi Gui Ze, GB13190-1991);

Guidelines for the Establishment and Development of Multilingual Thesauri (Duo Yu Zhong Xu Ci Biao Bian Zhi Gui Ze).

Those achievements have assured that various subject thesauri developed in subsequent years are consistent with the established international standard (ISO 2788). To further achieve compatibility among Chinese thesauri, a National Term Bank was also developed in the 1990s under the direction of the National Commission for Science and Technology (Yi, 1995). Since many of those subject thesauri were independently developed by different institutions, the terms contained in them were inconsistent. In the early 1990s the total number of the descriptors from all thesauri was close to 700,000 with a large number of overlaps. After nearly all of them being input into the database, there were about 250,000 preferred terms left. Designed as a compatibility center, its goal is to become a standardized vocabulary source for Chinese indexing languages. To facilitate automated switching of different thesauri in China, some experiments are also carried out in recent years (Zhang & Hou, 2000).

In order to meet the ever-increasing demand for information processing in sciences and technologies, the Institute of Scientific and Technical Information of China in late 1980s began to amend the section of natural sciences and technologies, the largest component of *CT*. The revised and enriched edition was published in 1991, which added 8221 new descriptors and deleted 5424 terms. For easy utilization, the size was reduced, the listing rearranged, and a vocabulary management system was also established.

A permuted index is a form of keyword-in-context (KWIC) index. The permuted index for natural sciences and technologies was added to the revised edition of the *CT* in 1996, and computer technologies have played a key role in its construction (Hou et al., 1998). It was compiled through the joint efforts of the Institute of Scientific and Technical Information of China and the Chinese Indexing Society, and published by the Science and Technology Press. In the form of a KWIC index with over 160,000 entries, it is especially useful in making accessible constituent terms that are not indexing terms and therefore do not occur anywhere else in the thesaurus.

Accomplishments & problems of the *CT*

As the most comprehensive thesaurus in Chinese language, the *CT* has played a key role in the development of Chinese information indexing and retrieval. It provides not only a rich vocabulary for subject indexing in various disciplines, but also a solid foundation for the subsequent construction of specialized thesauri of the past 20 years. Through its compilation process, Chinese information professionals have gained valuable experience in thesaurus construction by exploring different structures and techniques for both manual and automated indexing. It is fair to say that the publication of the *CT* has substantially promoted the use of subject indexing in Chinese information retrieval since 1980s.

The development of the *CT* clearly reflects the nature of the Chinese society in the late 20th century. Unlike in the Western countries where thesaurus compilation is usually conducted by a small group of individuals or private organizations, the Chinese central government agencies—the National Library and the Institute of Scientific and Technical Information of China—have played an active leadership role in the publication of the *CT*.

Thousands of library and information professionals from hundreds of government institutions contributed to the process. The development of the *CCT* and the *CLC* also follows the same pattern. In addition, the *CT* is basically designed according a systematic classification of disciplines, while in the US and Europe thesauri are usually developed for certain collections or projects. Furthermore, in order to reflect the character of the society and the history of the country many unique terms are included as descriptors in the *CT*, such as "the (Communist) Party's Guiding Principles for Economic Work" and "Cheng Ho's Expeditions to the Western Ocean." During the editorial process of the *CT*, every effort was made to ensure its compliance with the established international standards. The active involvement of central government authorities also contributed to its prompt adoption as a national standard in thesaurus compilation.

A valid concern regarding the effectiveness of the *CT* is related to its vocabulary control. Since it was designed to meet the needs of both comprehensive libraries as well as some specialized institutions of information services in China, this notion has landed the ambitious lexicon into a somewhat awkward position. As Zeng points out, a general thesaurus can neither reach as high a specialty nor cover as much special vocabulary as in a special one (Zeng, 1990, p. 94). As a collective effort, many individuals and institutions actively contributed to the compilation process. However in practice, it is very difficult to attain a balance between coverage and specificity. One way or another their individual marks are left in the final product of the *CT*. Related to this, another problem of the *CT* is its inconsistency. As mentioned earlier, the whole set was divided into three parts in 10 mighty volumes with separate listings of terms from social sciences, sciences and technologies. As a result, inconsistence exists between the two parts. Not only it is cumbersome to use, but also the integrity of the overall structure is compromised. In terms of quantity and depth of descriptors selected, discrepancy is also evident among different disciplines. On average, each major class has about 1872 terms. However in reality, the variation ranges from 7722 terms in a larger category to 274 from a smaller subject. Research indicated that for the past 10 years, only one-third of total terms have been used by the National Library of China in indexing recent Chinese publications. Comparing with other major thesauri in the world, the equivalence rate of terms in the *CT* is only 16.53%, far below the industry average of 50% (Ma & Hou, 1999). This is especially true in the areas of natural sciences and technologies. As a consequence, there are nearly 18,000 unrelated terms in the thesaurus, many of them are never or rarely used in practice. Therefore for a wider acceptance, a thorough examination and overhaul of the whole vocabulary is much needed if the *CT* is deemed to be a truly effectively tool in Chinese information index and retrieval.

As regards to the coordination of compound terms, more than 60% of all complex descriptors in the *CT* were pre-coordinated, above the recommended international standard of ISO 2788 (Hou & Ma, 1991). The high pre-coordination rate was due to certain unique features of flexible expressions in Chinese. Moreover, the development of the *CT* was to reflect the character of the society and the history of the country. In addition, ample consideration was given to the need of manual indexing. Yet during the course of thesaurus construction, a strict, uniform standard was not established in advance, to a certain degree it affected the quality of the pre-coordinated terms in the thesaurus. The inconsistent standards are also reflected in the construction of vocabulary indexes. While some classes have more than 2000 terms, others have only two descriptors. The depth of hierarchy varies from two to nine levels among disciplines. In addition, some duplication is present between the hierarchical index and the subject category index. The rough formation of the subject category index also reduces its functional effectiveness. In short, the lack of control and balance has curtailed the wide adoption of the *CT*.

The publication of the *CT* is still a remarkable achievement considering it was compiled before any national standards for thesaurus construction were developed. Examining further, many problems of the *CT* could be attributed to its lack of a supervising organization. The *CT* is a joint product between the Institute of Scientific and Technical Information of China and the National Library of China, with multiple resources pulled together from hundreds of other institutions. The long revision cycle without administrative control has severely limited its use, as no single authorized agency is fully responsible for the impending revision of the thesaurus. Evidently, to meet the growing needs in knowledge organization, index and retrieval, a dynamic management system must be in place for any future success in *CT* development.

CCT

China is a country with a long tradition in library classifications. In their professional practice,

Chinese librarians are very much familiar with the use of a library classification system. However this is not the case with the newly arrived *CT*. Many of them experienced difficulties in mastering subject indexing. Therefore since the early 1980s there has been an increasing demand for an effective tool to organize information that links both classification numbers and subject headings. For this reason a new effort was underway in 1986 to compile the *CCT*. Forty institutions participated in the project under the leadership of the National Library of China. This undertaking was also based on recent research trends in classified subject indexing and thesaurofacet (Hu, 1994; Yuan, 1997). According to Aitchison, a faceted thesaurus involves the development of a faceted classification as the systematic section, complemented by an alphabetical index or a full alphabetical thesaurus (Aitchison, Gomersall, & Ireland, 1969). The division of information between the systematic and the alphabetical sections was about equal, with the classification displaying the main hierarchies and related terms in a particular field, and the alphabetical section providing the information on scope notes, synonyms and related terms in other fields. However in China, since the *CLC* system as an enumerative classification has been widely adopted across the country, it would be impractical and unrealistic to have a new faceted classification developed for this purpose. Hence the compilation of the *CCT* took a modified approach in integrating both classification and subject indexing languages in China.

Published in 1994 in two parts, six volumes and 6215 pages, the *CCT* is again designed as a multifunctional indexing tool for Chinese publications (Editorial Committee of the Chinese Library Classification, 1994). Part I (volumes 1–2) consists of the classification number–subject descriptor corresponding table, and Part II (volumes 3–6) includes the subject descriptor–classification number corresponding table. With over 50,000 entries and 210,000 terms, it is not only a complete set of the *CLC*, but also a revised and enlarged *CT* (Liu, 1996). In the *CCT*, the corresponding table of subject headings–classification numbers functions as the tool for subject indexing. Through this two-way cross reference, the part of classification schedule can be taken for category index and hierarchical indexes of the *CT*, the part of thesaurus can be taken for class indexes of the *CLC*. The cumbersome subject category index is replaced by the simplified classification schedule from the *CLC*. This logical, systematic arrangement further improves the subject indexing function of the *CCT*. In addition, all terms in the alphabetical list, both social sciences and sciences and technologies, have been revamped, obsolete words deleted, new descriptors added, and arrangement modified for user convenience and efficiency.

To meet the challenge of automated information retrieval, a machine-readable copy of the *CCT* was also released, and the Zhongshan Library in Canton, Guangdong Province became the first library in China to successfully implement in its ZSLAIS regional public library system an integrated classification, subject indexing and retrieval system based on the *CLC* and *CCT* (Zhang, 1995). This experience shows that the *CCT* is suitable for both manual indexing as well as computerized retrieval. In an automated environment where post-coordinated indexing is fully utilized, a swift exchange between classification numbers and subject headings is realized, therefore users gain much better control of the system, not to mention the added benefits of easy vocabulary management.

As a large-scale, comprehensive indexing retrieval device, the *CCT* is to build a bridge between a library classification schedule and a traditional subject thesaurus. Based on the integration of the classification language and subject indexing, the *CCT* boasts a combined structure and dynamic functionality in Chinese information processing. The ultimate goals of the *CCT* are to promote the swift development of computer-based subject indexing and a national information retrieval network, and through this process, facilitate China's vigorous demands for knowledge and information in its ambitious modernization drive. The *CCT* is not to replace a classification system, but to incorporate the components of library classification and indexing into a smooth, integrated process of knowledge organization and retrieval. Through a few years of practical usage, it is fair to say that *CCT* has helped Chinese librarians and information professionals in embracing subject indexing languages. Nevertheless, its overall effectiveness and long-term impact are yet to be seen.

Conclusion

In China, the modern development of knowledge organization and information retrieval has come a long way. The landmark achievements in Chinese librarianship of the 20th century are the compilation of the *Chinese Library Classification* (*CLC*), the *Chinese Thesaurus* (*CT*) and the *Chinese Classified Thesaurus* (*CCT*). All three of them are the collective work of thousands of librarians and

information professionals. Looking back over the last 20 years, their concentrated efforts have gradually proceeded from classification language to subject indexing, and to integrated process of information organization and retrieval. The Chinese central government agencies such as the National Library and the Institute of Scientific and Technical Information of China have also played active leadership roles in those time-consuming, labor-intensive undertakings.

Despite many drawbacks, the *CT* and the *CCT* still are the largest, and most comprehensive indexing retrieval devices available for Chinese language. They have enhanced the standardization of Chinese retrieval language, and contributed greatly to the latest advancement in the field of knowledge organization and information processing in China. Based on these accomplishments, the challenge ahead is how to take advantage of new computer technologies, and construct a uniform national information network which will combine knowledge organization, subject indexing and retrieval into a seamless process for millions of Chinese language users. As we enter the exciting information age of the 21st century, what waits to be seen is how the *CT* and the *CCT* will evolve in the current revolution of Chinese information processing.

In the era of the Internet where computerized, natural language searching of full-text documents is trendy, a thesaurus can still remain as an important tool for information retrieval. The role of the thesaurus is changing, and the change is likely to represent an expansion rather than a contraction of the thesaurus's utility. The expansion may include not only a role of thesauri in performance enhancement in full-text systems, but also as a tool for use on websites, or for any or all of indexing, search statement expansion and visual organization. As Milstead states, thesauri are not obsolete, merely invisible (Milstead, 1995). Further research is needed to determine what modifications are most likely to make *CT* and *CCT* more useful in full-text systems, whether in indexing or searching. However such a topic is beyond the scope of this article.

Acknowledgements

The author wishes to express his appreciation to the two anonymous referees for their constructive criticism in the revision of this article.

References

Aitchison, J., Gilchrist, A., & Bawden, D. (2000). *Thesaurus construction and use: a practical manual*, 4th Edn. Chicago: Fitzroy Dearborn Publishers.

Aitchison, J., Gomersall, A., & Ireland, R. (1969). *Thesaurofacet: A thesaurus and faceted classification for engineering and related subjects*. Whetsone, Leicester, England: English Electronic Company Ltd.

Editorial Committee of the Chinese Library Classification (1994). *Chinese classified thesaurus* (Zhong Guo Fen Lei Zhu Ti Ci Biao). Beijing: Huayi Press.

Han Yu Zhi Ti Ci Biao (1980). *Chinese Thesaurus*. Beijing: Science and Technology Press (For clarification purpose, the standard Chinese pronunciation, Hanyu Pinyin, is given here for special Chinese terms like CT throughout the paper).

Hou, H. Q., & Ma, Z. H. (1991). *Introduction to subject thesaurus* (Zhu Ti Fa Dao Lu). Beijing: Peking University Press.

Hou, H. Q. et al. (1998). Computer-aided compilation of the permuted index of the Chinese thesaurus. *Journal of the Chinese Society for Scientific and Technical Information* (Qing Bao Xue Bao), *17*(4), 281–286.

Hu, M. (1994) A review of the theory of classification-subject integration thesaurus and its development. *Library & Information Work* (Tu Shu Qing Bao Gong Zuo), *1994*(5), 19–23.

Liu, X. S. (1996). The Compilation and use of integrated thesaurus combined classification with thesaurus in China. *Journal of the Hong Kong Library Association* (Xiang Gang Tu Shu Guan Xie Hui Xue Bao) *18*, 63–68.

Lu, Y. W. (1993). The activities in thesaurus construction for materials databases in China. In R. A. Strehlow, & S. E. Wright (Eds.) *Standardizing terminology for better communication: Practice, applied theory and results, ASTM STP 1166* (pp. 47–53). Philadelphia: American Society for Testing and Materials.

Ma, Z. H., & H. Q. Hou, (1999). *Introduction to classification and subject thesaurus* (Wen Xian Fen Lei Fa Zhu Ti Fa Dao Lu). Beijing: The National Library of China Press.

Milstead, J. L. (1995). Invisible thesauri: The year 2000. *Online and CD-ROM Review*, *19*(2), 93–94.

Yi, H. (1995). Indexing languages, new progress in China. *Knowledge Organization*, *22*(1), 30–32.

Yuan, X. Q. (1997). A comparative study of three integrated thesauri. *Journal of the Library Science in China* (Zhong Guo Tu Shu Guan Xue Bao), *23*(109), 27–32.

Zeng, L. (1986). An introduction to thesauri and classification systems in the People's Republic of China. *International Classification*, *13*(1), 24–28.

Zeng, L. (1990). Establishing a compatible general vocabulary in China: The capability. *International Classification*, *17*(2), 91–98.

Zhang, Q. (1995). On the structure and functions of the *Chinese classified subject thesaurus. Journal of the Library Science in China* (Zhong Guo Tu Shu Guan Xue Bao), *21*(99), 70–75.

Zhang, Q. Y. et al. (1996). Contemporary classification systems and thesauri in China. *Paper presented at the 62nd general conference of the international federation of library associations (IFLA) in Beijing, China.* In *IFLA'96 Booklet 4* (049-CLASS-3-E, August, 1996) (pp. 31–37).

Zhang, X. Y., & Hou, H. Q. (2000). Design of vocabulary switching system of thesauri. *Journal of the Chinese Society for Scientific and Technical Information* (Qing Bao Xue Bao), *19*(5), 451–457.

Classification and Categorization: A Difference that Makes a Difference

ELIN K. JACOB

ABSTRACT

EXAMINATION OF THE SYSTEMIC PROPERTIES AND FORMS of interaction that characterize classification and categorization reveals fundamental syntactic differences between the structure of classification systems and the structure of categorization systems. These distinctions lead to meaningful differences in the contexts within which information can be apprehended and influence the semantic information available to the individual. Structural and semantic differences between classification and categorization are differences that make a difference in the information environment by influencing the functional activities of an information system and by contributing to its constitution as an information environment.

INTRODUCTION

Many different and sometimes conflicting responses can be made to the question "What is information?" Floridi (in press) identifies three broad categories intended to elucidate the predominant approaches to understanding the ambiguous phenomenon called information: information as reality (or ecological information), information for reality (or instructional information), and information about reality (or semantic information). The approach adopted here is that information is "differences that make a difference" (Bateson, 1979, p. 99). It is an emergent property—the result of meaningful differences—inherently semantic and therefore about reality.

Analysis of the syntactic differences that distinguish systems of classification from systems of categorization can contribute to a philosophy of information (PI) because these distinctions portend significant consequences for the processes that contribute to what Floridi (2002) describes

Elin K. Jacob, School of Library and Information Science, Indiana University–Bloomington, 1320 E. 10th St., Bloomington, IN 47405–1801
LIBRARY TRENDS, Vol. 52, No. 3, Winter 2004, pp. 515–540
© 2004 The Board of Trustees, University of Illinois

as the "dynamics of information": "(i) *the constitution and modelling of information environments,* including their systemic properties, forms of interaction, internal developments etc.; (ii) *information life cycles,* i.e., the series of various stages in form and functional activity through which information can pass . . . and (iii) *computation,* both in the Turing-machine sense of *algorithmic processing* and in the wider sense of *information processing*" (p. 15. emphasis in original). Examination of the systemic properties and forms of interaction that characterize classification and categorization reveals fundamental differences in their respective organizational structures—differences that influence the functional activities of an information system and contribute to its constitution as an information environment.

The argument elaborated here is that fundamental syntactic distinctions exist between the structure of classification systems and the structure of categorization systems; that these distinctions lead to meaningful differences in the contexts within which information can be apprehended; and that these differences, in turn, influence the semantic information—the information about reality—that is available to the individual.

Information Systems

Shera (1960/1965) has observed that retrieval must be the focus of a theory of library and information science (LIS) and thus "the end toward which all our efforts are directed" (p. 136). Unfortunately, retrieval is too often viewed not as one component in an information system but as a self-contained and independent process. This emphasis on the end product—the retrieval of resources—tends to obscure the fact that effective retrieval depends on both the representation and the organization of a collection of information resources.

Soergel (1985) points out that, because information is used for problem-solving, information systems are developed and extended in response to the problems that confront society. Although this definition of information is not universally accepted, it is useful in understanding the complex set of processes that contribute to the ultimate effectiveness of an information system. Such a system identifies information resources that may be of use in addressing a particular problem; represents the attributes of resources that are relevant to the problem area; organizes these resource representations or the resources themselves for efficient access; and ultimately retrieves a set of resources in response to queries presented to the system by the individual. It would appear, then, that a more productive approach to the problem of retrieval would be to view an information system as a multidimensional whole comprised of several interrelated processes, including, at a minimum, collection development, representation, organization, and retrieval.

Retrieval is the final and therefore the most obvious of the processes that contribute to an information system. Because it is the only process in

which an individual actively participates, it is frequently the only process to which she gives serious consideration. When the individual is seeking information on a particular topic, her attention is focused on the set of resources retrieved by the information system. If these resources appear to be pertinent to the immediate problem, she may not give a second thought to the appropriateness of the terms used to query the information system. Nonetheless, it is the processes of selection, representation, and organization that provide the foundation without which information retrieval (IR) is less than effective, if not impossible. How resources are represented constrains the organizational structure(s) that can be imposed on a collection of information resources; the organizational structure of the collection dictates the search strategies that can be used for retrieval; and the representations themselves determine the set of resources that will be retrieved by the system.

Shera (1956/1965) affirmed the critical roles of representation and organization when he observed that effective retrieval requires an accord between the cognitive organization imposed on information by the individual and the formal organization imposed upon representations by the system. Shera's argument for accord between the individual and the retrieval system rests on three basic assumptions: that there are certain cognitive structures that can be identified and described; that it can be demonstrated that these structures are shared across individuals; and that identification of these shared structures will provide the basis for a theory of organization.

That cognitive accord can be achieved across individuals is a fundamental assumption of the shareability constraint proposed by Freyd (1983). She argued that the intent to communicate without loss of information causes the individual to modify her internal conceptual representations to reflect the cognitive organization assumed to be held by the other participant(s) in the communicative process. If participation in an intentional act of communication does promote normalization of conceptual representations across individuals, as Freyd (1983) argues, it follows that an intentional act of communication between the individual as natural intelligence and the information system would be subject to a similar shareability constraint. Assuming that the processes of representation, organization, and retrieval are necessarily interdependent, failure to address communication between the individual and the information system from the perspective of the system is a significant omission. Thus, an accounting of the dynamics of information should address the role of representation and organization in the creation and communication of meaningful information. More importantly, it should account for the semantic implications occasioned by differences in the forms of organization that can be used to structure an information system.

The need for effective communication between the information system and the individual points to five areas of research: (i) Is communication

between the information system and the individual influenced by the representation of resources? (ii) Does the organizational structure of the information system cause the individual to adjust her internal cognitive structures? (iii) Does the organization of resources contribute to the creation of a meaningful context for information? (iv) Is the meaning of information influenced by the organizational structure of the information system? and (v) What consequences follow from the different organizational structures that can be applied to a collection of information resources?

An understanding of the different forms of organizational structure and the implications that each holds for creating a meaningful context for information is foundational and must therefore precede any discussion of the role that representation and organization play in the dynamics of information. Accordingly, the focus here is on the ramifications of organizational structure for communication between the information system and the individual as natural intelligence. More specifically, the argument presented here addresses the fundamental structural and semantic differences between classification and categorization and how these differences make a difference in the information environment.

Categorization

Categorization is the process of dividing the world into groups of entities whose members are in some way similar to each other. Recognition of resemblance across entities and the subsequent aggregation of like entities into categories lead the individual to discover order in a complex environment. Without the ability to group entities based on perceived similarities, the individual's experience of any one entity would be totally unique and could not be extended to subsequent encounters with similar entities in the environment. Consider a situation in which each separate entity—each tree, each flower, or each drop of rain—was distinct from all other entities and carried its own unique set of defining characteristics. As Markman (1989) observes, the individual would not be able to handle the variety and complexity of her day-to-day interactions with the environment. By reducing the load on memory and facilitating the efficient storage and retrieval of information, categorization serves as the fundamental cognitive mechanism that simplifies the individual's experience of the environment.

Categorization divides the world of experience into groups or categories whose members share some perceptible similarity within a given context. That this context may vary and with it the composition of the category is the very basis for both the flexibility and the power of cognitive categorization. Zerubavel (1993) contends that the individual finds order and meaning in the environment by imposing boundaries—by splitting and lumping objects of experience so as to create distinct "islands of meaning" (p. 5). How an entity is categorized creates a context or conceptual frame that not only provides information about the entity but also shapes the in-

dividual's interaction with it. For example, the historic period known as the English Renaissance (1500–1650) is perceived as fundamentally different from the English Middle Ages even though England in the sixteenth century was, in many respects, quite similar to England in the fifteenth century. Splitting the sixteenth century from the fifteenth century by labeling them as belonging to two distinct historical periods focuses attention on the differences between them rather than on their similarities and provides the information that, in England, these differences were of greater import than differences between the fourteenth and fifteenth centuries.

Barsalou (1987) points out that this ability to manipulate the environment through the creation of categories allows the individual to forge new relationships and thus to create new information whose value exceeds the simple grouping of objects in the environment. He proposes that, because different features or properties are used to represent the same category at different times and in different contexts, the information associated with a particular category varies across individuals and across contexts. Thus the set of features associated with a category on any given occasion is composed of both context-dependent and context-independent information. Context-dependent information is relevant only within a particular context. For example, a high temperature of 50 degrees Fahrenheit might be described as cold on a summer day in southern Indiana, but warm or even hot on a winter day in the same locale. To say that it is cold outside conveys context-dependent information that is meaningful only in relation to the seasonal context. In contrast, context-independent information provides information about a category that is relevant across contexts. Even when used metaphorically, for example, the word "fire" connotes heat, light, and energy. The apparent instability of categories is therefore a reflection of the flexibility and the plasticity that are the power of the cognitive process of categorization and of the individual's ability to create and modify the informational content of a category as a function of immediate context, personal goals, or past experience.

The acquisition and transmission of information are dependent not only on the cognitive ability to create new categories—and thus new information—through the discovery of new patterns of similarity across entities, but also on the ability to capture information about these patterns through the medium of language. With the accumulation of more specialized knowledge and the creation of disciplinary domains, however, these categories and the relations between them have a tendency to become formalized (Jacob, 1994). The need to ensure that disciplinary knowledge is consistent across individuals and across time privileges the stability of reference provided by well-defined classes. As experientially-based categories evolve into well-defined, domain-specific classes that facilitate sharing of knowledge without loss of information, they lose their original flexibility and plasticity as well as the ability to respond to new patterns of similarity.

The Classical Theory of Categories

Until Rosch's publication in the 1970s of her seminal work on categories and categorization (Rosch, 1973, 1975), research in the area of categorization had focused on concept formation not as a process of creation but as a process of recognition. The world of experience was assumed to consist of a set of predetermined categories, each defined by a set of essential features represented by a category label; and all members of a given category were assumed to share a set of essential features that was identified by the category label and could be apprehended by all members of the linguistic community. Thus Hull (1920) wrote of the child's discovery of meaning in the word "dog" as the gradual recognition of a preexisting and invariant concept: "The 'dog' experiences appear at irregular intervals. . . . At length the time arrives when the child has a 'meaning' for the word dog. Upon examination this meaning is found to be actually a characteristic more or less common to all dogs and not common to cats, dolls and 'teddy-bears'" (Hull, 1920, pp. 5–6; cited in Brown, 1979, p. 188).

The presumption that a category is determined by a set of defining criteria is known as the "classical theory of categories." This is a simple but powerful theory that rests on three basic propositions (Smith & Medin, 1981; see also Taylor, 1989):

1. The intension of a category is a summary representation of an entire category of entities.
2. The essential features that comprise the intension of a category are individually necessary and jointly sufficient to determine membership within the category.
3. If a category (A) is nested within the superordinate category (B), the features that define category (B) are contained within the set of features that define category (A).

Proposition I states that the definition (*intension*) of a category is the union of the essential features that identify the membership (*extension*) of that category. Furthermore, because all members of a single category must share this set of essential features, each member is equally representative of the category as a whole. For this reason, the internal structure of a category is said to be ungraded, or without rank, because no member can be more typical or more representative of a category than any other member.

Proposition II states that, because each member of the category must exhibit all of the essential features that comprise the intension of the category, possession of the set of features that defines the category is sufficient to determine membership in the category. And, because there is a binary, either/or relationship that exists between an entity and a category such that an entity either is or is not a member of a particular category, the boundaries of categories are said to be fixed and rigid.

Proposition III identifies the inheritance relationship that exists be-

tween categories in a hierarchical structure: any member of a category that is a subset of a superordinate category must exhibit not only the set of essential features that determine membership in the subset but also the set of essential features that determine membership in any superordinate category within which the subset is nested.

In its most rudimentary form, categorization can be defined as the placement of entities in groups whose members bear some similarity to each other. Within the framework of the classical theory of categories, however, categorization is the process of systematically dividing up the world of experience into a formalized and potentially hierarchical structure of categories, each of which is defined by a unique set of essential feature(s). Because the intension of a category defines the set of essential features that each member of the category must exhibit, the classical theory maintains that intension equals extension—that membership within a particular category (extension) entails possession of the essential and defining character (intension) of the category. For example, if the intension of the category "bird" consists of the features "lays eggs," "has wings," "flies," and "builds nests in high places," every member of the category must exemplify the complete set of defining features. If an entity does not fly, it cannot be accorded membership in the category "bird" even if it does lay eggs, have wings, and build nests in high places. And, because all members of the category are defined by the same set of features, no one bird can be more typical or more representative of the category than any other bird. Thus, according to the classical theory, a parrot, a pigeon, and a puffin would be equally representative of the category "bird."

Brown (1979) observes that within the formalized and rigidly constrained ordering of reality established by the classical theory, category membership is absolute: ". . . any given thing is either in or out of the set" (p. 189). It is this stipulation that is the source of the classical theory's explanatory power: because it requires that intension equals extension—that membership in a category demonstrates possession of the set of essential features that define the category—the classical theory of categories would provide a simple yet elegant explanation for both the internal structure of cognitive representations and the semantic meanings of words.

Until recently, the classical theory of categories exemplified "the 'right way' to think about categories, concepts, and classifications" (Gardner, 1987, p. 340). But empirical research conducted over the past thirty years has challenged the validity of the assumptions on which this theory is founded. Critics of the classical theory have argued that the inability of subjects to identify the defining characteristics of an entity (Hampton, 1979; Rosch & Mervis, 1975) not only undermines the assumption that the set of essential features determining category membership is absolute but also calls into question the notion that these features are available to and can be specified by all members of a linguistic community. Demonstration of graded typi-

cality effects—the observation that subjects do judge certain members to be more representative of a category than others (McCloskey & Glucksberg, 1978; Rips, Shoben, & Smith, 1973; Rosch, 1973, 1975)—controverts the assumption that category structure is ungraded because all members are equally representative of the category. There is evidence, too, that subjects are able to rank both members and nonmembers of a category on a single continuum of representativeness. For example, Barsalou (1987) demonstrated that subjects could rank a robin, a pigeon, an ostrich, a butterfly, and a chair on a single continuum of representativeness for the category "bird"—a continuum extending from the most typical member of the category (robin) to the most atypical member (chair). The evidence for graded structure of categories points to the lack of fixed and determinate boundaries separating members of a category from nonmembers; and, buttressed as it is by demonstrations of category membership based on family resemblance (Rosch & Mervis, 1975), graded structure casts doubt on the classical assumption that there is an explicit inclusion/exclusion relationship between an entity and a category.

Classification

In LIS, the term "classification" is used to refer to three distinct but related concepts: a system of classes, ordered according to a predetermined set of principles and used to organize a set of entities; a group or class in a classification system; and the process of assigning entities to classes in a classification system. The focus here is on the first of these—on the classification system as a representational tool used to organize a collection of information resources—but a full appreciation of the implications of classification for information environments requires a basic understanding of the classification process itself.

Classification as process involves the orderly and systematic assignment of each entity to one and only one class within a system of mutually exclusive and nonoverlapping classes. This process is lawful and systematic: lawful because it is carried out in accordance with an established set of principles that governs the structure of classes and class relationships; and systematic because it mandates consistent application of these principles within the framework of a prescribed ordering of reality. The scheme itself is artificial and arbitrary: artificial because it is a tool created for the express purpose of establishing a meaningful organization; and arbitrary because the criteria used to define classes in the scheme reflect a single perspective of the domain to the exclusion of all other perspectives.

Taxonomic Classification.

Classification is perhaps best exemplified by the discipline of taxonomy. Broadly defined, taxonomy is the science of classification or, as Mayr (1982) defines it, "the theory and practice of delimiting kinds of organisms"

(p. 146). The objectives of taxonomic investigation are to provide an orderly and systematic organization of knowledge about the biological world; to identify the defining characteristics that distinguish a biological entity; and, based on those characteristics, to place the entity within a hierarchical ordering of mutually exclusive superordinate and subordinate classes in accordance with a set of established and widely accepted principles.

Taxonomic classification establishes stability of nomenclature through the aegis of a formalized and universally accepted language that facilitates transmission of knowledge across time and the barriers of natural language. Each class in the taxonomic scheme is given a unique name that is used to refer to all entities that display the complete set of features defining the class. And, because it is universally employed to identify all members of a given class, this label provides access to the accumulated knowledge about those entities, not as individuals but as members of a particular class. The taxonomic name establishes a relationship of equivalence between the set of features that define the class (its intension) and the set of entities that are members of the class (its extension). Using the taxonomic name, a member of a biological class is recognizable wherever it occurs, regardless of natural language or the local name(s) by which it may be known.

Through the inheritance of definitional criteria made possible by enforcing a principled structure of superordinate and subordinate classes, taxonomic classification also serves as an external cognitive scaffolding (Clark, 1997; Jacob 2001, 2002) that provides for the economical storage and retrieval of information about a class of entities. For example, the observation that Bleu is a poodle provides information about Bleu that is associated with the class "poodle." More importantly, however, it also provides information about Bleu that is available from the hierarchical structure within which the class "poodle" is nested—information associated with the superordinate classes dog, mammal, vertebrate, etc.

The essential observation, however, is that the practice of taxonomy is carried out within the arbitrary framework established by a set of universal principles. For example, while the naturalist Adanson, a contemporary of Linneaus, proposed a method for organizing botanical phenomena based on the identification of differences between individual specimens (Foucault, 1970), Linneaus advocated a systematic approach based on similarity of reproductive structure. For the naturalist following Linneaus's lead, any physical differences between two specimens not directly related to the process of reproduction would be irrelevant: for example, differences of leaf, stem, or root structure that might be used to distinguish between two plants would be ignored if the plants exhibited similar reproductive structures.

Taxonomic classification supports the efficient storage and retrieval of information about a class of entities, but reliance on a systematic approach such as that advocated by Linneaus constrains the information context by limiting the identification of knowledge-bearing associations to hierarchi-

cal relationships between classes. Furthermore, class definitions based on a single feature such as reproductive structure effectively reduce the amount of meaningful information that can be represented about each class in the taxonomy.

Classification Schemes.

A classification scheme is a set of mutually exclusive and nonoverlapping classes arranged within a hierarchical structure and reflecting a predetermined ordering of reality. Because a classification scheme mandates that an entity can be a member of one and only one class, it provides for communication of meaningful information through the systematic and principled ordering of classes. Furthermore, it establishes and enforces stability of reference by providing each class with a unique label that links individual members of the class to the class definition.

Shera (1951/1965) observes that, throughout history, attempts to classify knowledge have relied on four basic assumptions: universal order, unity of knowledge, similarity of class members, and intrinsic essence. The assumption of universal order posits an immutable conception of reality that serves as a unifying framework for all knowledge. The assumption of unity of knowledge presupposes that past, present, and future knowledge can be represented within a single, inclusive hierarchy of superordinate and subordinate classes. The assumption of similarity of class members holds that a class can be defined by a set of essential features and that these features are shared by all members of the class and distinguish that class from all other classes in the structure. And the assumption of intrinsic essence maintains that there is a set of individually necessary and jointly sufficient features that is intrinsic to all members of a class and that these features constitute the essence of the class.

With the possible exception of universal order, Shera's exposition of the assumptions that support efforts to organize knowledge can be interpreted in terms of the three propositions that constitute the classical theory of categories: the assertion that a category is defined by a summary representation (Proposition I) is a statement of the essential similarity of class members; the assertion that a category is defined by a set of essential features (Proposition II) is a statement of the intrinsic essence of a class; and the assertion that defining features are inherited in a hierarchical structure of categories (Proposition III) is a statement of the unity of all knowledge. It is instructive that, although the classical theory of categories is unable to account for the variability and flexibility of cognitive categorization, it does provide an elegant accounting of the fundamental assumptions on which classification schemes have historically been constructed.

Bibliographic Classification Schemes.

Traditionally, bibliographic classifications have been deductive, top-down schemes that enumerate a set of mutually exclusive classes. An enu-

merative classification scheme begins with a universe of knowledge and a theory of organization or set of principles that establishes the conceptual structure of the scheme. Whether the universe encompasses all knowledge or is limited to a specific domain, construction of the scheme involves the logical process of division and subdivision of the original universe such that each class, or each level of classes in the structure, is differentiated by a particular characteristic or property (e.g., the property "color" or "shape"). The result is a hierarchical structure of generic (genus/species) relationships wherein each subordinate class is, theoretically, a true species of the superordinate within which it is nested.

Faceted (analytico-synthetic) classification systems are inductive, bottom-up schemes generated through a process of analysis and synthesis. Construction of the faceted structure begins with analysis of a universe of knowledge to identify the individual elements—properties and features—of the universe. These elements are then organized into mutually exclusive groups on the basis of conceptual similarity, and these groups are, in turn, arranged in successively larger groupings to form facets (aspects) that can be used to represent entities in the universe. In this way, meaningful relationships are established not only between the elements in a group but between the groups themselves. The result is not a classification scheme but a controlled vocabulary of concepts and their associated labels that can be used, in association with a notation and a prescribed citation order, to synthesize the classes that will populate the classification scheme. A faceted vocabulary for classifying cars might include mutually exclusive facets for "color" (red, blue, black), "body style" (sedan, convertible, minivan), and "transmission" (manual, automatic). Following the citation order *body style—transmission—color*, classes would be constructed by selecting a single value, or isolate, from each facet. Examples of the classes that could be constructed in this faceted scheme would be convertible—manual—red and minivan—automatic—blue.

Because a faceted classification scheme adheres to a fixed citation order during the construction of individual classes, the resulting structure, like an enumerative scheme, is necessarily hierarchical. In fact, it is the hierarchical nature of bibliographic classification schemes that allows for the arrangement of physical resources on the shelves of the library. "Reading" a classification scheme involves moving down the hierarchy, from superordinate to subordinate and from left to right, to generate a series of relationships between classes that can be translated into the linear order of the library shelf. It is just this linear structure that Ranganathan captured in the notion of APUPA (or Alien-Penumbral-Umbral-Penumbral-Alien). The umbral class (U) represents the focal topic; penumbral classes (P) are those most closely related to the focal topic; and alien classes (A) are those removed from and therefore unrelated to the focal topic. When the individual reviews a collection of resources arranged in classified order, she

generally begins with the most relevant class or focal topic (U); moving either to the right or the left, she progresses from resources on the focal topic through closely related materials (P) to those resources which are unrelated (A). In this fashion, the linearity inherent in the hierarchical structure of a classification scheme is used to create a meaningful context by bringing into proximity those classes within the hierarchical structure which are theoretically most closely related.

Linearity is, in fact, the first of seven properties that Shera (1953/1965) identifies as characteristic of a bibliographic classification scheme: linearity; inclusivity of all knowledge within the classification's universe; well-defined, specific, and meaningful class labels; an arrangement of classes that establishes meaningful relationships between them; distinctions between classes that are meaningful; a mutually exclusive and nonoverlapping class structure; and an infinite hospitality that can accommodate every entity in the bibliographic universe. Each of these properties contributes to Shera's definition of a bibliographic classification scheme as

> a list of terms which are specifically and significantly different each from the other, capable of describing the subject content of [resources], inclusive of all knowledge, infinitely hospitable, in an arrangement that is linear, unique, and meaningful, and which when applied to [resources], usually, though not necessarily, through the medium of a notation, results in their arrangement on the shelves according to the logical principles that inhere in the schematism. (Shera, 1953/1965, p. 99)

In other words, a bibliographic classification establishes a controlled vocabulary in the form of a set of uniquely labeled classes that serve both to define and to organize the intellectual content of a collection of resources. Furthermore, this vocabulary determines the conceptual boundaries of the scheme's universe by including only the knowledge that is relevant within the immediate universe. The resulting arrangement is meaningful precisely because it constitutes a principled context for information—a context shaped by class definitions, by information-bearing, hierarchical relationships and by meaningful distinctions between classes and, by extension, between the concepts that those classes represent.

Classification as a Disciplinary Language.

A classificatory structure frequently inheres in a disciplinary language when it is used to establish a specific conceptual context that both defines and organizes the domain of investigation (Foucault, 1970; Jacob, 1994). The language serves to prescribe the boundaries of the domain; to determine both the subject matter of the domain and the relationships that obtain between phenomena of investigation; to legitimize specific concepts and methodologies; to ensure effective transmission of knowledge by stabilizing the vocabulary; and to foster a domain-specific perspective or dis-

ciplinary episteme. Because a disciplinary language reflects the underlying classificatory structure of the domain, the meaning of any class term can only be apprehended within the conceptual context established by the classificatory structure.

The Difference between Classification and Categorization

Although there are obvious similarities between classification and categorization, the differences between them have significant implications for the constitution of an information environment. Failure to distinguish between these two systems of organization appears to stem from the misconception that they are, in fact, synonymous—a misconception that may be reinforced by the fact that both are mechanisms for organizing information.

The literature on categorization is riddled with passages where the terms "classification" and "categorization" are used indiscriminately to refer to the same process. Rosch et al. (1976) provides an illustrative example of how these two terms are used indiscriminately:

> . . . one purpose of *categorization* is to reduce the infinite differences among stimuli to behaviorally and cognitively usable proportions. It is to the organism's advantage not to differentiate one stimulus from others when that differentiation is irrelevant for the purposes at hand. The basic level of *classification*, the primary level at which cuts are made in the environment, appears to result from the combination of these two principles; the basic *categorization* is the most general and inclusive level at which categories can delineate real-world correlational structures. (Rosch et al., 1976, p. 384. Emphasis added)

This lack of distinction between *category/categorization* and *class/classification* is frequently compounded by the use of *concept* as yet another synonym for *category* (e.g., Gardner, 1987, p. 340). Unfortunately, this terminological imprecision obscures the fact that researchers are actually dealing with two similar but nonetheless distinct approaches to organization.

Although systems of classification and categorization are both mechanisms for establishing order through the grouping of related phenomena, fundamental differences between them influence how that order is effected—differences that do make a difference in the information contexts established by each of these systems. While traditional classification is rigorous in that it mandates that an entity either is or is not a member of a particular class, the process of categorization is flexible and creative and draws nonbinding associations between entities—associations that are based not on a set of predetermined principles but on the simple recognition of similarities that exist across a set of entities. Classification divides a universe of entities into an arbitrary system of mutually exclusive and nonoverlapping classes that are arranged within the conceptual context established by

a set of established principles. The fact that neither the context nor the composition of these classes varies is the basis for the stability of reference provided by a system of classification. In contrast, categorization divides the world of experience into groups or categories whose members bear some immediate similarity within a given context. That this context may vary—and with it the composition of the category—is the basis for both the flexibility and the power of cognitive categorization (Jacob, 1992).

Figure 1 identifies six systemic properties that serve as a starting point for comparing systems of classification and categorization: (i) process, (ii) boundaries, (iii) membership, (iv) criteria for assignment, (v) typicality, and (vi) structure.

(i) The process of classification involves systematic arrangement of classes of entities based on analysis of the set of individually necessary and jointly sufficient characteristics that defines each class. In contrast, the process of categorization is generally unsystematic but inherently creative in that it need not rely on predetermined definitions but is able to respond to similarity assessments based on immediate context, personal goals, or individual experience.

Figure 1. Comparison of Categorization and Classification.

Categorization	Classification
Process	
Creative synthesis of entities based on context or perceived similarity	Systematic arrangement of entities based on analysis of necessary and sufficient characteristics
Boundaries	
Because membership in any group is non-binding, boundaries are "fuzzy"	Because classes are mutually-exclusive and non-overlapping, boundaries are fixed
Membership	
Flexible: category membership is based on generalized knowledge and/or immediate context	Rigorous: an entity either *is* or *is not* a member of a particular class based on the intension of a class
Criteria for Assignment	
Criteria both context-dependent and context-independent	Criteria are predetermined guidelines or principles
Typicality	
Individual members can be rank-ordered by typicality (graded structure)	All members are equally representative (ungraded structure)
Structure	
Clusters of entities; may form hierarchical structure	Hierarchical structure of fixed classes

(ii) Systems of classification and categorization are also distinguished by the boundaries imposed on groupings. Because the classes in a classification system are rigidly circumscribed by the intension of the class and further constrained by the requirement that they be mutually exclusive and nonoverlapping, boundaries between classes are fixed, determinate, and persistent. In a categorization system, however, membership of an entity in any one category is nonbinding and does not prohibit membership in any other category. Thus the membership of any two or more categories in a system of categorization may overlap or vary across time in response to changing contexts. This is possible because category boundaries are not simply fuzzy but are, in fact, mutable and potentially fluid.

(iii) and (iv) Membership and criteria for assignment are two closely related characteristics that distinguish systems of classification from systems of categorization. In a classification system, criteria for class assignment—the set of necessary and sufficient features that constitutes the intension of a class—are governed by principles that establish the conceptual framework of the system. Membership in a class is rigorous in that it is determined by the intension of the class: an entity either is or is not a member of any class in the system. More importantly, however, membership in a class is absolute simply because an entity can belong to one and only one class. In contrast, the criteria for category assignment employed by a system of categorization are potentially variable, allowing the membership of a category to respond to the demands of the context in which it is used. In this way, the membership of a category may vary across time based on the combination of context-dependent and context-independent information that is used to define category membership.

Differences in the criteria for assignment emphasize an important distinction between classification and categorization. In systems of classification, class assignment relies on definitions that are "idealizations" or "theoretical abstractions" (Barsalou, 1987) to determine class membership. In systems of categorization, however, category assignment is flexible and dynamic, reflecting the ability of the individual to modify category definitions in response to variations in the immediate environment. Thus Barsalou argues that

> . . . the concepts that theorists "discover" for categories may never be identical to an actual concept that someone uses. Instead, they may be analytic fictions that are central tendencies or idealizations of actual concepts. Although such theoretical abstractions may be useful or sufficient for certain scientific purposes, it may be more fruitful and accurate to describe the variety of concepts that can be constructed for a category and to understand the process that generates them. (Barsalou, 1987, p. 120)

(v) Typicality is closely related to the characteristics of membership and criteria for assignment. However, typicality is potentially ambiguous: on the

one hand, typicality is used as an indication of the individual's assessment as to how representative a member is of its particular class or category; and, on the other hand, it is used as a reflection of the assumptions regarding membership and membership criteria that govern a system of classification or categorization. Because empirical research indicates that subjects are capable of ranking members according to typicality even when working with well-defined, either/or classes such as *odd number* or *even number* (Armstrong, Gleitman, & Gleitman, 1983), attempting to distinguish between classification and categorization on the basis of an individual's typicality judgments would be an exercise in futility. In contrast, systemic assumptions governing membership do provide an important point of distinction between classification and categorization.

In a system of classification, all members of a class must display the full set of essential features prescribed by the class definition (see Proposition I of the classical theory). It follows, then, that all members are assumed to be equal and therefore equally representative of the class. For this reason, the internal structure of a class is said to be ungraded because no entity can be a "better" member of the class than any other member. However, in a system of categorization, there is no assumption of equality of membership. The fact that individuals can identify particular members as more typical of a category reflects the dynamic nature of category definitions and the corresponding variability of category membership as a reflection of immediate context. The internal structure of a category is said to be ungraded because it is possible to rank category members as to how typical or representative they are of the category as a whole.

(vi) Structure is perhaps the single most important characteristic that can be used to discriminate between systems of classification and categorization because it is influenced by distinctions based on process, boundaries, membership, and criteria for assignment. A classification system is generally a hierarchical structure of well-defined, mutually exclusive, and nonoverlapping classes nested in a series of superordinate-subordinate or genus-species relationships. The structure of a classification system provides a powerful cognitive tool—an external scaffolding (Clark, 1997; Jacob 2001, 2002)—that minimizes the cognitive load on the individual by embedding information about reality through the organization of classes within the system. For example, because an entity either is or is not a member of a particular class in a system of classification, it provides for determination of class membership as a relatively simple pattern-matching or pattern-completing activity. At a more complex level, the structure of the classification system establishes information-bearing relationships between classes: vertical relationships between superordinate and subordinate classes that are subject to the mechanism of inheritance illustrated above in the example of the poodle Bleu; and lateral relationships between coordinate classes that occur at the same level in the hierarchy and, when taken together, consti-

tute the immediately superordinate class within which they are nested. In this fashion, the structure of a classification system serves as a medium for the accumulation, storage, and communication of information associated with each class in the structure; and, by capitalizing on the hierarchical and lateral relationships between classes, it minimizes the information that must be stored with each class and reduces the load on memory.

In contrast, the structure of a categorization system consists of variable clusters of entities that may or may not be organized in a hierarchical structure. Because categories are not constrained by a requirement for mutual-exclusivity, membership in one category does not prohibit membership in any other category. More importantly, however, the very plasticity that is the creative power of categories may actually prohibit the use of categorization as a persistent information structure. The potentially transitory and overlapping nature of categories provides that any relationships established between categories are themselves mutable. Thus a system of categorization creates a conceptual framework whose meaning may be short-lived and ephemeral—a conceptual framework that cannot function as cognitive scaffolding and whose ability to serve as a medium for the accumulation, storage, and communication of information is limited.

Ordering, Grouping, and Organization

A system for ordering (Jacob & Loehrlein, 2003) provides access to resources by arranging them in some recognizable order. Typically, these systems will employ alphanumeric or chronological sequences because these arrangements generate syntactic patterns that are familiar to a majority of individuals. Although such a system is intended to support access to known items, it may appear to create groupings of similar resources (e.g., all individuals with the last name *Smith* or alumni who graduated in the year 2000), but the imposition of sequential order is nonetheless a purely syntactic device that cannot create meaningful relationships either between individual entities or between groups of entities.

In contrast, a system of organization (Jacob & Loehrlein, 2003) is a unified structure that establishes a network of relationships among the classes or categories that comprise the system. These relationships are meaningful and information-bearing because they specify principled connections between two or more groups within the same system. Thus, with a single possible exception, classification systems are systems of organization because they provide for the conceptual arrangement of a set of mutually exclusive and nonoverlapping classes within a systematic structure of hierarchical, genus-species relationships.

The exception is a constitutive classification (Jacob, Mostafa, & Quiroga, 1997) consisting of a set of mutually exclusive classes that comprise the totality of a given universe but lack nested, superordinate-subordinate relationships. For example, the classes freshman, sophomore, junior, and

senior comprise the universe of college undergraduates. These classes appear to evince a hierarchical ordering (e.g., from freshman to senior), but they fail to demonstrate meaningful, information-bearing relationships: although a senior may be assumed to have been a junior at some point in time, the class junior is not a true species of its purported superordinate senior. Thus a constitutive classification does not qualify as a system of organization because, even though it is comprised of a set of mutually exclusive and nonoverlapping classes that constitute the totality of a particular universe, it fails to establish meaningful relationships between its constituent classes. It is interesting, too, that neither a hierarchical nor a constitutive classification can serve as a system for ordering: because the distinctions between classes are conceptual, the classes cannot conform to a recognizable, syntactic pattern of arrangement. Furthermore, both hierarchical and constitutive systems of classification require an index or other auxiliary mechanism to support access, whether to unique resources or to individual classes in the structure.

A system of categorization may or may not be a system of organization. Although a categorization system groups entities on the basis of similarity, the example of a constitutive classification demonstrates that the simple identification of a set of categories without the establishment of meaningful, information-bearing relationships does not constitute a system of organization. But, even though a categorization system does not indicate meaningful relationships, it is not a system for ordering: the simple fact of grouping entities into categories does not support access. Because categorization reflects conceptual distinctions between groups of entities, it, too, requires an auxiliary mechanism to provide access, whether to individual categories or to unique category members.

If a system of categorization does not impose a systematic, syntactic order on its member categories and if it does not establish meaningful relationships between categories, then it is simply a mechanism for grouping. For example, dividing the items on a shopping list into categories defined by place of purchase (e.g., grocery store, gas station, and five-and-dime store) is a mechanism for grouping that simplifies the individual's interaction with her environment but neither creates meaningful relationships between categories nor imposes any recognizable order on them. A constitutive classification is also an example of a simple mechanism for grouping: in this case, for dividing a universe of entities into a set of well-defined and mutually exclusive groups without the identification of any meaningful relationships among them.

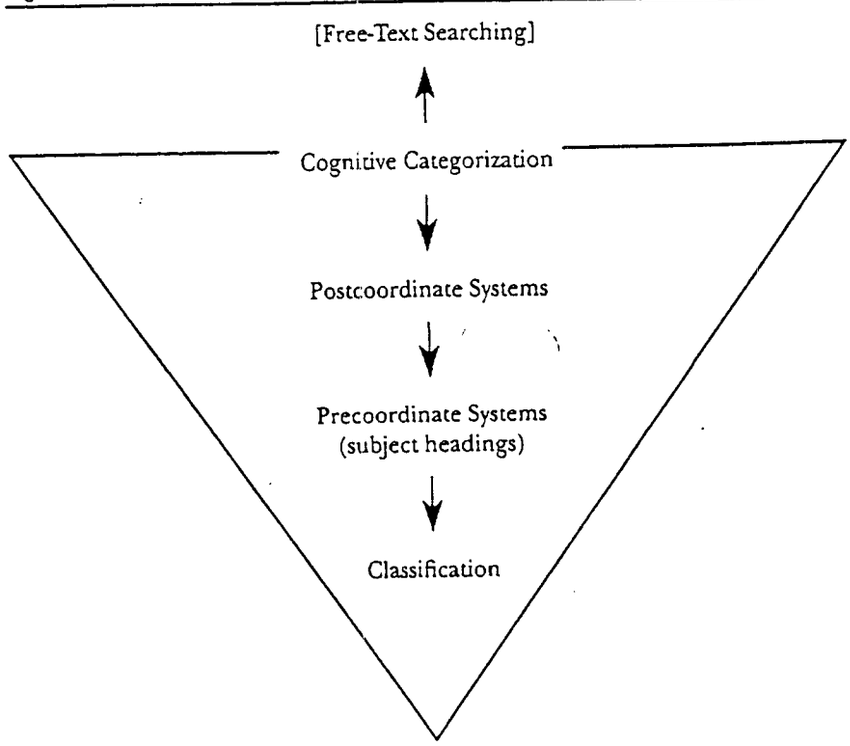
Implications of Structure

The functional role of structure in the creation and enhancement of information contexts can be addressed through analysis of four general approaches to the organization and retrieval of resources: free-text search-

ing, postcoordinate indexing, precoordinate indexing, and classification (see Figure 2). Although cognitive categorization serves as the baseline for this analysis, it is removed from consideration as a system of organization, not because it lacks semantic foundation or relational structure, but because, contrary to the arguments proffered by Shera (1956/1965), the organization imposed on cognitive categories is so dynamic and responsive to changes in context that it cannot establish persistent, knowledge-bearing relationships between categories.

Of the four general approaches to organization, free-text searching is the least constrained. Although it shares with systems of classification the creation of mutually exclusive, nonoverlapping, and rigidly bounded classes whose membership is constrained by an explicit criterion of assignment (i.e., the alphanumeric search string used to query the system), free-text searching lacks an established set of principles that governs the structure of classes and class relationships. It can be described as a system of categorization in the very broadest sense, but it is, at best, a very elementary mechanism for grouping. Even as a grouping mechanism, however, it has two significant flaws. In the first place, the basis for grouping is purely syntactic: because the criterion for group assignment involves the simple match-

Figure 2. Systems of Organization.



ing of alphanumeric strings, groups produced by this process share a superficial similarity without deeper semantic implications. In the second place, the process of free-text grouping is binary in that it generates only two groups of entities—those that match the query string and those that do not. However, because free-text searching lacks a semantic base, it cannot support meaningful distinctions between these two classes, and, because it exemplifies the very simplest of structures (i.e., two antonymous classes), a free-text retrieval system cannot contribute to an information environment that will support or enhance the value of system output through the establishment of meaningful context.

Unlike free-text searching, postcoordinate systems, precoordinate systems, and classification systems are all indexing systems in that each involves the assignment to a resource of one or more descriptors intended to represent the intellectual content of that resource. These descriptors are usually drawn from a controlled vocabulary or indexing language that normalizes the vocabulary used in representation and retrieval by creating an indexical, one-for-one correspondence between a descriptor and the concept to which it points. The indexing language also provides for communication between the system and the individual by specifying the set of authorized terms or subject strings that can be used to pose search queries to the system. Although a descriptor may be a class label, a subject heading or a single term or phrase, depending on the nature of the system, each descriptor serves to identify or describe the intellectual content of a group of resources. Unlike an access point in a system for ordering that supports the retrieval of a unique entity, a descriptor is a surrogate for (or a pointer to) the intellectual content shared by a group of resources. Indeed, indexing, like categorization, would be impossible if every resource were to be treated as a unique entity.

In the progression from postcoordinate indexing systems through precoordinate indexing systems to systems of classification, organizational structure becomes increasingly more constrained (see Figure 2). It is appropriate, then, to begin this analysis with classification, the most highly constrained of these three systems, and to work back through the less constrained systems toward the baseline of cognitive categorization.

Theoretically, a classificatory structure epitomizes a system of organization because it creates a principled structure of well-defined classes that are linked by a system of hierarchical, genus-species relationships. Although practice does not always adhere to theory in the development of classification schemes, classification is nonetheless the most rigid of organizational systems because its structure of mutually exclusive and nonoverlapping classes mandates an absolute relationship between a resource and its class: each resource may be assigned to one and only one class in the structure. Thus the process of classification is inherently systematic because it is gov-

erned by a set of principles that serves as a persistent conceptual framework for the creation of meaningful, structural relationships between classes.

Although the well-defined structure of a classification system provides for the creation of meaningful, information-bearing relationships between classes—relationships that facilitate the use of classification as an external cognitive scaffolding—it places powerful limitations on communication between the individual and the information system. In an information system whose class structure is predetermined, the retrieval set returned for any query posed to the system is necessarily limited to the membership of a single class. Thus the structure of the classification system constrains the questions that can be presented to the system by prescribing the set of possible answers before a query has actually been posed. Within a classificatory structure, then, communication is one-way—from the system to the individual—and the individual must rely on her understanding of or intuitions about the structural relationships between classes in order to interact with the system in an effective and meaningful way.

Information systems are identified as precoordinate when the categories or classes that comprise the system are either assigned or built by the indexer at the time of indexing. A classification system is obviously a precoordinate system because its classes are either established by the classificationist during scheme generation or built by the classifier at the time of class assignment using a faceted vocabulary and a fixed citation order. A subject heading system is also a precoordinate system but it is generally less constrained—and less constraining—than a classification system. Whereas classification mandates assignment of a resource to one and only one class, a precoordinate system of subject headings does not require individual groups to be mutually exclusive. Rather, subject heading systems allow for the assignment of multiple descriptors to a single resource, thereby providing multiple access points for each entity rather than the single access point (the unique class label) prescribed by a classification system.

Because it does not demand a well-defined and absolute relationship between a resource and a subject heading—because it does not require that the groups of entities associated with individual subject headings will necessarily be mutually exclusive—a precoordinate subject heading system is, in fact, a system of categorization. Categories formed by the subject heading system are not rigidly bounded but frequently overlap, with individual members spilling over into penumbral and even alien categories. Although allowing multiple descriptors for a single resource provides for greater variability in the range of resources that can be retrieved with a single query, the questions that can be posed to the information system are nonetheless limited, as they are in a system of classification, by the authorized set of subject heading strings that comprise the system. And, as with a classification system, the retrieval set generated in response to a query is determined

by the indexer: the assignment of subject headings as descriptors not only constrains the questions that can be posed to the system but serves to establish the specific set of resources that can be retrieved in response to each query posed to the system.

Unlike the systematic and principled structure of a classification system, the structure of a subject heading system is frequently unprincipled, unsystematic and polyhierarchical. And, unlike the relationships established between well-defined and mutually exclusive classes in a classification, any relationships created between the categories of a subject heading system cannot be assumed to be either meaningful or information-bearing. An example from *Subject Headings for Schools and Public Libraries* (Fountain, 2001) illustrates the lack of knowledge-bearing relationships that characterizes many subject heading systems. The heading "Rats as carriers of disease" combines two broader concepts: "rats" and "disease." Although it is obvious that "Rats as carriers of disease" is somehow related to both rats and disease, this heading is neither a kind of "Rat" nor a kind of "Disease." Because the specific value of any relationship that might link this heading to its broader concepts is unidentified, the relationship must be supplied by the individual if the heading is to be linked in a meaningful way to other concepts in the subject heading system.

Although subject heading systems appear to create relationships between headings, these relationships are often descriptive, idiosyncratic, and, sometimes, potentially meaningless. For example, the *Library of Congress Subject Headings* (Library of Congress. Cataloging Policy and Support Office, Library Services, 2002) identifies the subject heading "Humanities" as the broader term for the heading "Philosophy." It then proceeds to list "Humanism" as the broader term for "Humanities" and "Philosophy" as the broader term for "Humanism." Thus the supposed nesting structure is circular: "Philosophy" > "Humanities" > "Humanism" > "Philosophy." Obviously, the absence of either a well-defined indexing language or principled and meaningful relationships between subject headings undermines the ability of the system to establish a context that can contribute to the apprehension of information.

As with classification, communication between the individual and a subject heading system tends to be one-way—from the system to the individual—but the unprincipled structure of many subject heading systems and the general lack of a prescriptive conceptual framework that can support information-bearing relations undermines the potential for meaningful communication between the user and the system. This is an important distinction between subject heading systems and the more structured classification system that can be explained, in part, as a difference between the processes of identification and predication. Classification involves a process of identification (or definition) in that it asserts a meaningful, one-for-one relationship between an entity and its class, but a precoordinate system of

subject headings involves a process of predication (or description) that allows for multiple assertions to be ascribed to a single resource. While a system based on predication demonstrates greater creativity, flexibility, and hospitality than the well-defined structure of a system based on identification, the very rigidity of the latter actually supports the creation and persistence of information-bearing relationships that are simply not possible in the looser structure of the former.

Precoordinate systems constrain communication between the individual and the system through the establishment of a finite collection of class labels or subject headings that serve as the complete set of possible search queries and predetermine the composition of retrieval sets. In contrast, postcoordinate systems predetermine neither the queries nor the retrieval sets but allow the individual to build her own category definitions that can be presented to the system as search queries at the time of retrieval. Descriptors representing the intellectual content of a resource are assigned by the indexer at the time of indexing. During retrieval, the individual builds her own search categories by combining descriptors with Boolean logic.

By allowing the individual to generate her own queries, the postcoordinate system supports a more interactive form of communication between searcher and system. In most postcoordinate systems, descriptors are assigned from a controlled vocabulary. In others, however, communication between the individual and the information system is complicated by the fact that the indexing language does not exist as a controlled vocabulary but is extracted by the indexer from terms occurring in the resource being indexed. Generally, however, the generation of category definitions as postcoordinate search queries is limited only by the set of individual terms that comprise the indexing language. Although the resources that participate in a retrieval set are determined by the indexer's assignment of descriptors, communication between the system and the individual is greatly enhanced by her ability to create her own queries that will capture her immediate information need.

Unfortunately, however, the flexibility of category generation, like the process of cognitive categorization, goes hand-in-hand with the absence of meaningful relationships. As with any free-text information system, posing a query to a postcoordinate system simply divides the collection into two groups: the set of resources whose assigned descriptors match the search query and the remaining resources whose descriptors do not match the query. Obviously, postcoordinate systems, like free-text systems, are simply mechanisms for grouping, not systems of organization. Unlike free-text systems, however, the basis for grouping in a postcoordinate system is semantic, not syntactic. Although the postcoordinate system is simply matching strings, the indexer imposes a certain level of conceptual control by assigning simple descriptors from an indexing language that establishes an indexical, one-for-one relationship between a descriptor and its referent.

The individual is empowered to create unique and potentially idiosyncratic search categories precisely because the system itself does not establish any but the simplest categories—those defined by the individual descriptors assigned by an indexer. Because the system fails to establish a principled framework that provides for the establishment of information-bearing relationships between categories, the postcoordinate system can neither create nor contribute to an information context precisely because there is no persistent structure that could support meaningful relationships between categories.

Conclusion

This very preliminary review of the properties and features of the different approaches to organizing, ordering, or simply grouping information resources has barely scratched the surface in addressing structural distinctions between systems of classification and systems of categorization and how these distinctions affect interaction with the system as an information environment.

For example, at a very superficial level, the strength of classification is its ability to establish relationships between classes that are stable and meaningful. But the rigidity of structure that supports these relationships has its corresponding disadvantages. In particular, traditional classification systems are context-independent: because the relationships established by classification are invariant and persist across time and space, these systems are resilient to the context of use and severely constrain the individual's ability to communicate with the system in a meaningful and productive manner. In contrast, systems of categorization, and especially postcoordinate systems, are highly responsive to—even dependent on—the immediate context. The utility of these systems as information environments depends ultimately on provisions for effective communication with the individual. But the responsiveness and flexibility of the postcoordinate system effectively prohibit the establishment of meaningful relationships because categories are created by the individual, not the system, and are thus fleeting and ephemeral.

It is important for philosophers, theoreticians, and developers to work toward a more in-depth and comprehensive understanding of how the structure of an information system contributes to the establishment of semantic context; how different forms of organization support communication between the searcher and the system; and how concrete organizational structures and specific types of relationships contribute to the production of meaningful information environments. The search for adequate explanations of these issues will ultimately contribute to a deeper understanding of the "dynamics of information" (Floridi, 2002) and the implications that the structure of information systems holds for the composition of and interaction with information environments.

ACKNOWLEDGMENTS

I would like to thank Aaron Loehrlein for his thoughtful reading of and commentary on preliminary drafts of this paper and for the many conversations that contributed so richly to its theoretical content. I would also like to thank Ken Herold for his very careful and insightful reading of the final draft. This exploration of the role of structure in the generation of a semantically meaningful information environment is in its nascent stages, and I would like to thank Ken for the opportunity to develop these ideas for presentation in this venue.

REFERENCES

Armstrong, S. L., Gleitman, L. R., & Gleitman, H. (1983). What some concepts might not be. *Cognition, 13,* 263–308.

Barsalou, L. W. (1987). The instability of graded structure: Implications for the nature of concepts. In U. Neisser (Ed.), *Concepts and conceptual development: The ecological and intellectual factors in categorization* (pp. 101–140). Cambridge: Cambridge University Press.

Bateson, G. (1979). *Mind and nature: A necessary unity.* New York: Dutton.

Brown, R. (1979). Cognitive categories. In R. A. Kasschau & C. N. Cofer (Eds.), *Psychology's second century: Enduring issues* (pp. 188–217). New York: Praeger.

Clark, A. (1997). *Being there: Putting brain, body, and world together again.* Cambridge, MA: MIT Press.

Floridi, L. (2002). What is the philosophy of information? *Metaphilosophy, 33*(1/2). Retrieved August 20, 2003, from http://www.wolfson.ox.ac.uk/~floridi/pdf/wipi.pdf.

Floridi, L. (in press). Open problems in the philosophy of information. *Metaphilosophy.* Retrieved August 20, 2003, from http://www.wolfson.ox.ac.uk/~floridi/pdf/oppi.pdf.

Foucault, M. (1970). *The order of things: An archaeology of the human sciences.* New York: Vintage Books.

Fountain, J. F. (2001). *Subject headings for school and public libraries: An LCSH/Sears companion* (3rd ed.). Englewood, CO: Libraries Unlimited.

Freyd, J. J. (1983). Shareability: The social psychology of epistemology. *Cognitive Science, 7,* 191–210.

Gardner, H. (1987). *The mind's new science: A history of the cognitive revolution.* New York: Basic Books.

Hampton, J. A. (1979). Polymorphous concepts in semantic memory. *Journal of Verbal Learning and Verbal Behavior, 18,* 441–461.

Hull, C. L. (1920). Quantitative aspects of the evolution of concepts. *Psychological Monographs, 28,* Whole No. 123. [Cited in Brown, 1979]

Jacob, E. K. (1992). Classification and categorization: Drawing the line. In B. H. Kwasnik & R. Fidel (Eds.), *Advances in classification research, Vol. 2. Proceedings of the 2nd ASIS SIG/CR Classification Workshop: Held at the 54th ASIS Annual Meeting, Washington, DC, October 27–31, 1991* (pp. 67–83). Medford, NJ: Learned Information.

Jacob, E. K. (1994). Classification and crossdisciplinary communication: Breaching the boundaries imposed by classificatory structure. In H. Albrechtson & S. Oernager (Eds.), *Knowledge organization and quality management: Advances in knowledge organization, vol. 4* (pp. 101–108). Frankfurt/Main: Indeks Verlag.

Jacob, E. K. (2001). The everyday world of work: Two approaches to the investigation of classification in context. *Journal of Documentation, 57*(1), 76–99.

Jacob, E. K. (2002). Augmenting human capabilities: Classification as cognitive scaffolding. In M. J. López-Huertas & F. J. Muñoz-Fernández (Eds.), *Challenges in knowledge representation and organization for the 21st century. Integration of knowledge across boundaries: Proceedings of the Seventh International ISKO Conference, 10–13 July 2002, Granada, Spain* (pp. 38–44). Würzburg, Germany: Ergon Verlag.

Jacob, E. K., & Loehrlein, A. (2003). What ontologies are not: A [draft] theoretical framework for the analysis of representational systems. Paper presented in the SLIS Colloquium

Series, School of Library and Information Science, Indiana University-Bloomington. Retrieved August 20, 2003, from http://ella.slis.indiana.edu/~aloehrle/repsys.ppt.

Jacob, E. K., Mostafa, J., & Quiroga, L. M. (1997). An approach to the evaluation of automatically generated classification schemes. In P. Solomon (Ed.), *Advances in classification research, Vol. 7. Proceedings of the 7th ASIS SIG/CR Classification Workshop: Held at the 59th ASIS Annual Meeting, Baltimore, MD, October 20, 1996* (pp. 78–98). Medford, NJ: Information Today.

Library of Congress. Cataloging Policy and Support Office. (2002). *Library of Congress Subject Headings* (25th ed.). Washington, DC: Library of Congress, Cataloging Distribution Service.

Markman, E. M. (1989). *Categorization and naming in children: Problems of induction.* Cambridge, MA: MIT Press.

Mayr, E. (1982). *The growth of biological thought: Diversity, evolution, and inheritance.* Cambridge, MA: Harvard University Press.

McCloskey, M. E., & Glucksberg, S. (1978). Natural categories: Well defined or fuzzy sets? *Memory and Cognition, 6*(4), 462–472.

Rips, L. J., Shoben, E. J., & Smith, E. E. (1973). Semantic distance and the verification of semantic relations. *Journal of Verbal Learning and Verbal Behavior, 12*, 1–20.

Rosch, E. (1973). Natural categories. *Cognitive Psychology, 4*(3), 328–350.

Rosch, E. (1975). Cognitive representations of semantic categories. *Journal of Experimental Psychology: General, 104*, 192–233.

Rosch, E., & Mervis, C. B. (1975). Family resemblances: Studies in the internal structure of categories. *Cognitive Psychology, 7*(4), 573–605.

Rosch, E., Mervis, C. B., Gray, W., Johnson, D., & Boyes-Braem, P. (1976). Basic objects in natural categories. *Cognitive Psychology, 8*(4), 382–439.

Shera, J. H. (1951/1965). Classification as the basis of bibliographic organization. In *Libraries and the organization of knowledge* (pp. 77–96). Hamden, CT: Archon. (Reprinted from *Bibliographic organization*, pp. 72–93, 1951, Chicago, IL: University of Chicago Press)

Shera, J. H. (1953/1965). Classification: Current functions and applications to the subject analysis of library materials. In *Libraries and the organization of knowledge* (pp. 97–111). Hamden, CT: Archon. (Reprinted from *The subject analysis of library materials*, pp. 29–42, by M. F. Tauber, Ed., 1953, New York: Columbia School of Library Science)

Shera, J. H. (1956/1965). Putting knowledge to work. In *Libraries and the organization of knowledge* (pp. 51–62). Hamden, CT: Archon. (Reprinted from *Special Libraries, 47*, pp. 322–326, 1956)

Shera, J. H. (1960/1965). What lies ahead in classification. In *Libraries and the organization of knowledge* (pp. 129–142). Hamden, CT: Archon. (Reprinted from *Proceedings of the Allerton Park Institute*, pp. 116–128, 1960, Champaign: University of Illinois Bookstore)

Smith, E., & Medin, D. (1981). *Categories and concepts.* Cambridge, MA: Harvard University Press.

Soergel, D. (1985). *Organizing information: Principles of data base and retrieval systems.* Orlando, FL: Academic Press.

Taylor, J. R. (1989). *Linguistic categorization: Prototypes in linguistic theory.* Oxford: Clarendon Press.

Zerubavel, E. (1993). *The fine line: Making distinctions in everyday life.* Chicago, IL: University of Chicago Press.

Interoperable Summary Description Model Using Dublin Core

Do-Nyun Kim and Young-Won Song
LG Electronics, Information Technology Lab.,
16 Woomyeon-Dong, Secho-Gu, Seoul 137-724, Korea
dnkim@LG-Elite.com and ysong@LG-Elite.com

Abstract

This paper proposes an interoperable metadata model generating summary description for multimedia content using Dublin Core (DC). The motive is based on the fundamental concepts such as (1) Description information about the multimedia content is essential in multimedia content access, search and retrieval process (2) the existing metadata are too complicated to use in applications such as e-cataloguing and browsing of e-commerce. As an approach to solve the problem, summary description that may be optimally minimal descriptive elements set derived from existing metadata schemes (full descriptions) is described in this paper. The proposed summary description generator model is achieved using thesaurus approach built on the basis of DC and any existing various metadata schemes.

Keywords: summary description, interoperable, agent, thesaurus, Dublin Core, cataloguing

1 Introduction

Recently, the demand and appetite of multimedia contents consumer or end-users is increasing rapidly in virtue of the development and deployment of internet technologies. As this volume of demand increases, the importance and usefulness of metadata is emphasized and drives many metadata frameworks to be developed in order to build an infrastructure for the easy and interoperable delivery and consumption of multimedia content. However, unfortunately, there is no flexible, uniform, interoperable or integrated metadata framework that can be applied and reaching a worldwide consensus in the multimedia consumption value chain exists. In addition, some description metadata frameworks or schemes are too complex to be understood and implemented with reasonable cost. As a result, metadata users or sometimes, consumers tend to be overwhelmed by their complexities. As an approach to solve the current problems, summary description using DC that is considered as an essential and optimally minimal description set is proposed in this paper. Basically, this paper proposes to employ thesaurus approach in building the interoperable summary description model of DC element set derived from existing a variety of full multimedia description sets[1][2]. Section 2 addresses basic approaches to build the proposed model. Based on the basic approaches, section 3 describes the proposed interoperable summary description model in detail. Section 4 shows an example of summary description composed of DC element set derived from MPEG (Moving Picture Experts Group)-7 MDS (Multimedia Description Schemes) description set. In section 5, the validity of the proposed model is concluded regarding the practical e-commerce applications.

2 Basic approaches

In order to establish a summary description that is interoperable, optimally minimal, dynamic and maintainable, three fundamental approaches are employed and explained in this section as follow.

- Multiple-to-one mapping mechanism

The proposed model employs multiple-to-one mechanism that extracts DC elements set (one) from a variety of existing description schemes (multiple) containing full description information about multimedia contents. Since the purpose of this multiple-to-one mechanism is to obtain a DC element set from a variety of metadata element sets like a 'metadata filter', it is expected that the interoperability can be achieved. In a practical viewpoint, the 'DC metadata filter' is very useful in the e-commerce applications such as electronic cataloguing and browsing. One of the challenging tasks is to implement such mechanism that is able to generate *a* summary description automatically through summarization mechanism. One of the

©2001 National Institute of Informatics

reasonable methods is to use thesaurus approach.

- System-selected vs. User-selected elements

In the multimedia content search, cataloguing, browsing and retrieval, the optimized and minimal set of descriptive elements plays a key role in the process. However, since it is very difficult to define the optimally minimized set and to provide an interoperable metadata framework, the existing metadata frameworks may contain enormous number of elements as possible in order to overcome the practical applicability. As a fundamental approach, the DC is used. Also, in DC society, in order to provide more general metadata scheme, hierarchical qualifiers are recommended and used. However, still, there is an interoperability issue remained with the existing metadata frameworks such as MPEG-7 MDS, EPICS and etc[5][6][7]. This paper proposes to employ DC as a summary description. If end-users wish to investigate the more detail metadata, the full description can be accessed as a hierarchical manner. Since this DC summary description is extracted from the full description set, it is called as system-selected elements in this paper. In order to provide a flexible model with the end-user like user interaction, user defined elements (or user-selected elements) can be declared in addition to DC in the model such as event reporting that is defined in the MPEG-21 multimedia framework. In the viewpoint of interoperability, system-selected elements are efficient. User-selected elements are, however, more useful in the viewpoint of user interaction. Those two approaches are employed by summarization process described in section 3.

- Dynamic generation by users' demand

Summary description is a sort of filtered version of it's the original full description by nature. It is required to be generated by users' demand on time. This requirement is expected to prevent users from wasting efforts and cost. In terms of storage efficiency, summary description is not necessary to be stored permanently, because it is related to a specific application. Hence, the model is required that summary description interacts with the full description in a dynamic manner in response to end-user's demand. In such case, the summary description maintenance issue such as summary metadata consistency with the full description and version control in terms of metadata updatability is required to be solved. As a plausible approach to the dynamic summary description generation, software agent technique is proposed. In this situation, summarization process takes advantage of agents such as automatic notification and automatic change tracking.

Figure 1 shows the relationship among resource, summary description, (full) description, metadata of summary description (or Admin Core, Administrative Container Metadata, A-Core, or DescriptionMetadata in MDS), and metadata of (full) description (or A-Core). The full description describes resource such as audio, video clip, image, textual document or software. The metadata of summary or full description such as A-Core describes simply the summary or full description (for example, who, where, when the description is prepared). Hence, the summary description is different from metadata of description such as A-Core [3][4].

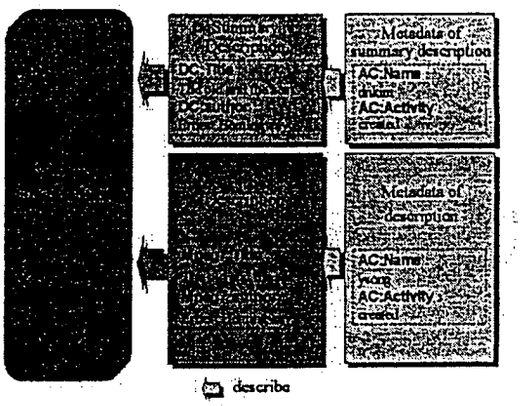


Figure 1. The relationship among resource, description, summary description, metadata of summary description and metadata of description

3 The proposed summary description model

This section describes the summary description model built based on the fundamental approaches addressed in section 2.

- Using Dublin Core

As a summary description, DC elements set is used. In the initial access, search and browsing procedure about the metadata, summary

description can play a major role in the perspective of end-users (In general they want simple, fast and automatic process). If users are interested in the corresponding content, users may request more detail description based on full description. Therefore, summary description needs not to be complicated and DC that has a wide consensus among e-commerce communities is sufficient to describe essential minimal description.

In case any search fails using DC or user-selected elements, or even if search results hit end-users' interests but they want to find detail description data, end-users may access the full description for more detail in a hierarchical manner as shown in Figure 2.

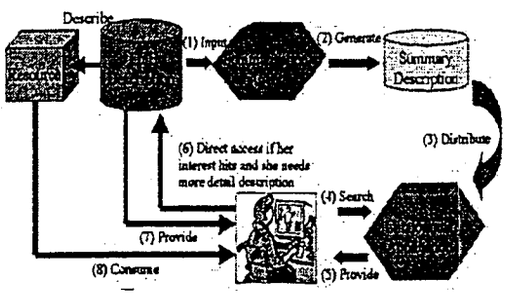


Figure 2. The sequence of actions using summary and full description in a hierarchical manner

- Using thesaurus

As described in section 2, the summarization process reflects many-to-one mechanism. In order to reduce ambiguous translation, the agent embedded in the summarization process utilizes thesaurus or dictionary – sometimes called controlled vocabulary. Thesaurus includes synonyms and antonyms. When elements of multiple and diverse descriptions are translated into those of summary description that is considered to be matched with DC elements set, the summarization process needs synonyms, antonyms and acronyms. Well-organized thesaurus has a significant influence on the performance of summarization agents.

Figure 3 describes the summarization process including thesaurus regarding to existing full description[5][6][7].

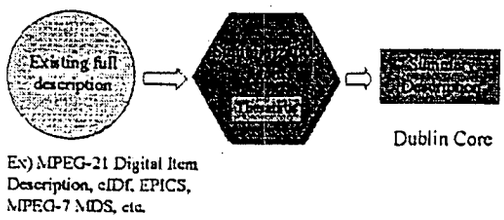


Figure 3. The input and output of the summarization process

Figure 4 shows a generic flow diagram of the summarization process.

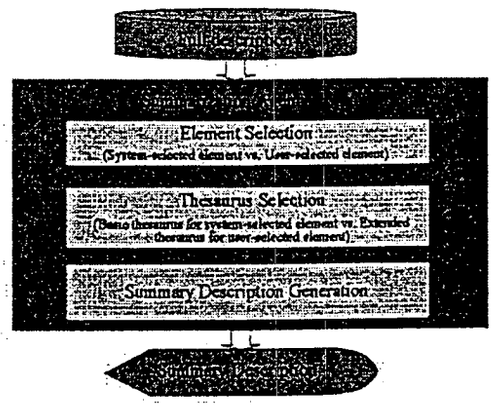


Figure 4. Flow diagram of summarization process

Figure 5 shows a block diagram that summary description is used in electronic cataloguing application.

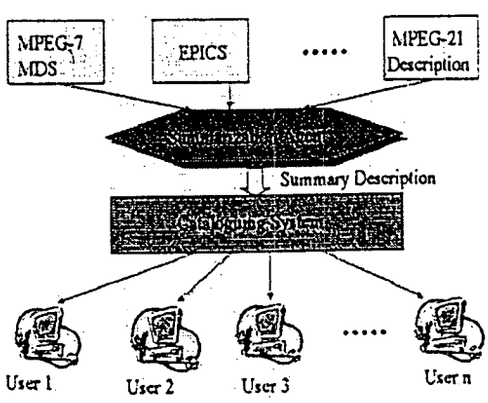


Figure 5. Summary description usage in e-cataloguing application

4 Example of Thesaurus

Figure 6 shows a part of thesaurus example that is obtained using MPEG-7 MDS and DC that can be performed by summarization agent [1][2][4].

Title – [element][A name given to the resource]
- ⇨ <MPEG-7 MDS><Creation><Title type="main">
- ⇨ Alternative - [qualifier][Any form of the title used as a substitute or alternative to the formal title of the resource]
  - ■ <MPEG-7 MDS> <Creation> <Title type="alternative">
  - ■ <MPEG-7 MDS> <Creation> <Title type="secondary">
  - ■ <MPEG-7 MDS> <Creation> <Title type="popular">

Description – [element][An account of the content of the resource]
- ⇨ Abstract – [qualifier][The topic of the content of the resource
  - ■ <MPEG-7 MDS> <Creation> <Abstract> <FreeTextAnnotation>
  - ■ <MPEG-7 MDS> <Creation> <Abstract> <StructuredAnnotation>

Creator – [element][An entity primarily responsible for making the content of the resource]
- ⇨ <MPEG-7 MDS> <Creation> <Creator>

Subject – [element][The topic of the content of the resource]
- ⇨ <MPEG-7 MDS> <Classification> <Subject>

Language – [element][A language of the intellectual content of the resource]
- ⇨ <MPEG-7 MDS> <Classification> <Language>

Coverage – [element][The extent or scope of the content of the resource]
- ⇨ <MPEG-7 MDS> <UsageInformation> <Availability>
- ⇨ Spatial – [qualifier] [Spatial characteristics of the intellectual content of the resource]
  - ■ <MPEG-7 MDS> <UsageInformation> <Availability><OriginPlace>
- ⇨ Temporal – [qualifier][Temporal characteristics of the intellectual content of the resource]
  - ■ <MPEG-7 MDS> <UsageInformation> <Availability> <AvailabilityPeriod>

Publisher – [element][An entity responsible for making the resource available]
- ■ <MPEG-7 MDS> <UsageInformation> <Availability><Distributor>

Figure 6. An example of thesaurus for MPEG-7 MDS

Note. In case of Title, it is a DC element. The meaning is a titled name of the corresponding resource. The matching element in MPEG-7 MDS is <Title> element under <Creation>. If <Creation> is not specified before the <Title>, summarization process confuses with the <Title> of <PersonName> specified also in MPEG-7 MDS that indicates the titles of an individual, such as honorifics.

5 Conclusions

The conceptual model of interoperable summary description derived from the existing metadata frameworks or schemes is proposed. The validity and usefulness of the model is described. The summary description example obtained using MPEG-7 MDS shows the validity of the proposed model employing DC and thesaurus approaches. This summary description is expected to be meaningful in e-cataloguing and browsing applications of e-commerce.

6 References

[1] Dublin Core Metadata Initiative, *Dublin Core Metadata Element Set, Version 1.1: Reference Description*, July 2, 1999.

[2] Dublin Core Metadata Initiative, *Dublin Core Qualifiers*, July 11, 2000.

[3] R. Iannella and D. Campbell, *The A-Core: Metadata about Content Metadata*, June 30, 1999.

[4] ISO/IEC JTC1/SC29/WG11 N4040: *Information technology – Multimedia Content Description Interface– Part5: Multimedia Description Schemes, CD, Singapore*, March 2001.

[5] Godfrey Rust and Mark Bide, *The <indecs> metadata framework- Principles, model and data dictionary*, WP1a-006-2.0, June 2000.

[6] Carl Lagoze, Clifford A. Lynch and Ron Daniel, Jr.: *A Container Architecture for Aggregating Sets of Metadata*, June 28, 1996.

[7] R.V.Guha and Tim Bray, *Meta Content Framework Using XML*, http://www.w3.org/TR/ NOTE-MCF-XML -970624, submitted to W3C 6 June, 2000.

Available online at www.sciencedirect.com
SCIENCE DIRECT
ELSEVIER
Computers & Operations Research 31 (2004) 2387–2404
computers & operations research
www.elsevier.com/locate/dsw

Improving search results with data mining in a thematic search engine

M. Caramia[a,*], G. Felici[b], A. Pezzoli[c]

[a] *Istituto per le Applicazioni del Calcolo IAC-CNR, Viale del Policlinico, 137, 00161 Roma, Italy*
[b] *Istituto di Analisi dei Sistemi ed Informatica IASI-CNR, Viale Manzoni 30, 00185 Roma, Italy*
[c] *Istituto di Analisi dei Sistemi ed Informatica IASI-CNR, Viale Manzoni 30, 00185 Roma, Italy*

Abstract

The problem of obtaining relevant results in web searching has been tackled with several approaches. Although very effective techniques are currently used by the most popular search engines when no a priori knowledge on the user's desires beside the search keywords is available, in different settings it is conceivable to design search methods that operate on a thematic database of web pages that refer to a common body of knowledge or to specific sets of users. We have considered such premises to design and develop a search method that deploys data mining and optimization techniques to provide a more significant and restricted set of pages as the final result of a user search. We adopt a vectorization method based on *search context* and *user profile* to apply clustering techniques that are then refined by a specially designed genetic algorithm. In this paper we describe the method, its implementation, the algorithms applied, and discuss some experiments that has been run on test sets of web pages.
© 2003 Elsevier Ltd. All rights reserved.

Keywords: Search engines; Web mining; Clustering; Genetic algorithms

1. Introduction

The recent improvements of search engines technologies have made available to the internet users an enormous amount of knowledge, that can be accessed in many different ways. The most popular search engines on the market are now providing search facilities for databases containing billions of web pages, where queries are executed instantly. These extraordinary results have been possible thanks to evoluted indexing systems and very powerful and efficient software and hardware components. Although, many researchers still question whether it is possible to construct search

* Corresponding author. Fax: +39-06-4404306.
E-mail addresses: caramia@iac.rm.cnr.it (M. Caramia), felici@iasi.rm.cnr.it (G. Felici).

0305-0548/$ - see front matter © 2003 Elsevier Ltd. All rights reserved.
doi:10.1016/S0305-0548(03)00194-1

tools that are able to provide more significant results to the users. The focus is thus switching from *quantity* (maintaining and indexing large databases of web pages and quickly select pages matching some criterion) to *quality* (identifying pages with a high quality for the user). Such a trend is motivated by the natural evolution of the internet users, that are now more selective in their choice of the search tool and may be willing to pay the price of providing extra feedback to the system and wait more time in order to have their queries better matched. In this framework, relevant work has been done in [1,2], and several have considered the use of data mining and optimization techniques, that are often referred to as *web mining*.

This paper presents a method for improving standard search results in a thematic search engine, where the documents and the pages made available are restricted to a finite number of topics, as well as the users are considered to belong to a finite number of user profiles. The method uses clustering techniques to identify some structure in the starting set of pages, and then a metaheuristic that tries to optimize the query results based on the results of the clustering.

The use of these techniques is not new in web mining applications. There has been extensive research on how to improve retrieval results by employing clustering techniques. In several studies the strategy was to build a clustering of the entire document collection and then match the query to the cluster centroids (see for example [3]). More recently, clustering has been used for helping the user in browsing a collection of documents [4] or in organizing the results returned by a search engine [5,6], or by a metasearch engine [7] in response to a user query. In [8] document clustering has also been used to automatically generate hierarchical clusters of documents. A different approach uses already existing document taxonomy as clusters to produce an effective document classifier for unclassified documents [9]. As discussed in [10], the use of clustering in information retrieval (IR) is based mostly on the *cluster hypothesis*: "closely associated documents tend to be relevant to the same request". Several researchers have shown that the cluster hypothesis also holds in a retrieved set of documents [11], but they do not study how the clustering structure may help a user in finding relevant results more quickly.

Metaheuristics, and more precisely, genetic algorithms, have been implemented in IR by several researchers and the results indicate that these algorithms could be efficient. In [12] Gordon presents a genetic algorithm based approach to improve document indexing. In that approach, the initial population is represented by a collection of documents judged relevant by a user, which is then evolved through generations and converges to an optimal population with a set of keywords which best describe the documents. In [13] the same author adopts a similar approach to document clustering, where a genetic algorithm is used to adapt subject descriptions so that documents become more effective in matching relevant queries.

In [14] the authors apply genetic algorithm in information retrieval in order to improve search queries that produce better results according to users' preferences. In [15] Chen uses genetic algorithms to optimize keywords that were used to suggest relevant documents. Also in [14–19] several approaches based on genetic algorithms to enhance the query description are discussed. In [20] a genetic algorithm is deployed to find an optimal set of documents that best match the user's need. In [21] a method for extracting keywords from documents and assigning them weights is proposed. Both documents and queries are represented as vectors in which the components correspond to significant keywords extracted from documents. A genetic algorithm is then used for assigning weights to the keywords; such weights are then used to form a query vector through which relevant documents may be detected. In [22] a prototype agent for collecting and filtering information from

the Internet according to a user profile is described. This intelligent agent includes a genetic algorithm which collects and evaluates new HTML documents from the Web, using information included in examples provided by the user, and returns the user with pages that have high score.

In this paper, yet another way of using genetic algorithms in search engines is proposed. We first use clustering to identify, in the set of pages resulting from a simple query, subsets that are homogeneous with respect to a vectorization based on *context* or *profile*; then we construct a number of small and potentially good subsets of pages, extracting from each cluster the pages with higher scores. Operating on these subsets with a genetic algorithm, we identify the subset that has the property of combining a good overall score with a high internal dissimilarity. This provides the user with few non-duplicated pages that represent more correctly the structure of the initial set of pages.

The outline of the paper is as follows. In Section 2 we provide a general description of the method, illustrating its basic components and its characteristics. Sections 3–5 treat in more detail each of the steps that define the method: page vectorization, page clustering and the implementation of the genetic algorithm, respectively. In Section 6 we summarize some experiments that were conducted with the proposed method, based both on randomly generated instances and on real data. Some concluding remarks are given in Section 7.

2. Description of the method

Let P be a set of web pages, and indicate with $p \in P$ a page in that set. Now assume that P is the result of a standard query to a database of pages, and thus represents a set of pages that satisfy some conditions expressed by the user (for example, that satisfy a boolean expression on some search keywords). Each page $p \in P$ is associated with a score, based on the query that generated P, that would determine the order by which the pages are presented to the user who submits the query. The role of this ordering is crucial for the quality of the search: in fact, if the dimension of P is relevant, the probability that the user considers a page p strongly decreases as the position of p in the order increases.

This may lead to two major drawbacks: (a) the pages in the first positions may be very similar (or even equal) to each other; (b) pages that do not have a very high score but are representative of some aspect of set P may appear in a very low position in the ordering, with a negligible chance of being looked at by the user. Our method tries to overcome both drawbacks, focusing on the determination of a small set of pages, selected from the initial set P, that, beside containing pages with a high score, also are different from each other and are chosen from different regions of some space where the pages are represented.

One of the conditions needed to apply our approach is the availability of additional information from the user, who indicates a search context or a user profile, where:

- the search context is a general topic to which the search is referred to, that is not necessarily linked with the search keywords that generated the set P; it may be viewed as an item of a topic catalog or directory as the ones that are frequently provided by the last generation of search engines (i.e., catalogs);
- the user profile is a subjective identification of the user, that may be either provided directly by choosing amongst a set of predefined profiles, or extracted from the pages that have been visited more recently by that user.

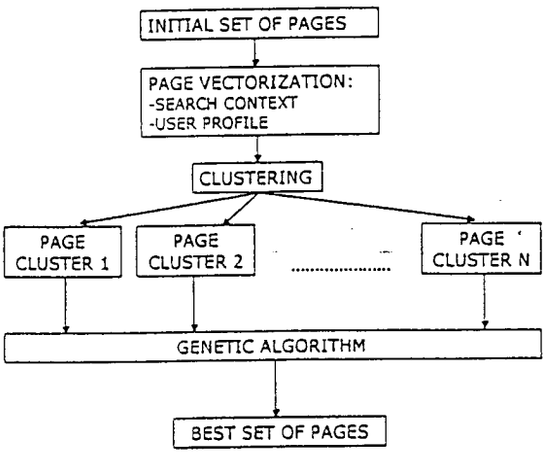


Fig. 1. The scheme of the method.

The basic idea of the method is to use the information conveyed by the search context or the user profile to analyze the structure of set P and determine in it an optimal small subset that better represents all the information available. This is done in three steps, that we sketch below and describe in the following sections.

First, the search context and the user profile are used to extract a finite set of significant words or page characteristics that is used to create, from all pages in P, a vector of characteristics (page vectorization). Such vectorization represents a particular way of "looking" at the page, specific of each context/profile.

Second, the vectorized pages are analyzed by a *clustering algorithm* that partitions them into subsets of similar pages. This induces a two-dimensional ordering on the pages, as each page p can now be ordered according to the original score within its cluster. At this point the objective is to provide the user with a reduced list that takes into account the structure identified by the clusters and the original score function. This is done in the third step, where the pages that have higher score in each cluster are selected to compose an initial population that is then analyzed by a *genetic algorithm*. The initial population is formed by subsets of pages, called chromosomes in the genetic algorithms terminology, extracted in turn from the clusters, starting from the pages with higher score; this way the initial population is composed of pages with high score that have a high probability of being different from each other. The genetic algorithm operates using a *fitness function* that combines the score with a measure of the similarity of the pages in the same chromosome. Once the genetic algorithm has been run, the best chromosome constitutes the subset of pages to be returned to the user.

Fig. 1 depicts the overall scheme of how the system works.

The quality of the final subset of pages depends, amongst other things, on the way the genetic algorithm is implemented; that is, how the fitness of each chromosome is measured and how chromosomes are selected and combined in each iteration. The main idea is that, when properly designed, the genetic algorithm can generally determine chromosomes that are sufficiently heterogeneous and whose pages have good values for the original score. We devote Section 5 to these details.

One may question whether the described method can be run on-line in a search engine as the standard execution of a user's query. We believe that with proper bounds on the parameters and a proper choice of the algorithms deployed, the computational effort can be dealt with and that only few seconds may be added to the overall search process. Moreover, the computation power available on commercial personal computers has and will increase significantly in the future.

Another issue that may rise related to the applicability of this method is that a high level of interaction with the user is required, as a search context has to be selected or be identified into a given profile. But the average skill of the internet user increased very rapidly, together with the need for precise and effective search methods. For this reason users should be willing to provide the additional information needed as the price for better results.

3. Page vectorization

The first step of the method is the representation of each page that has been acquired by a vector of finite dimension m, where each component represents a measure of some characteristic of the page (page vectorization). In the most plain setting, each component of the vector is the number of occurrences of a particular word; one may also consider other measurable characteristics that are not specifically linked with the words that are contained in the page, such as the presence of pictures, tables, banners and so on. As mentioned above, the vectorization is based on one context, or one profile, chosen by the user. One may then assume that, for each of the contextes/profiles that have been implemented in the search engine, a list of words that are relevant to that context/profile is available and a related vectorization of the page is stored. Clearly, many refinements to this simple approach may, and should, be considered. First of all, one should not consider simple "words", but rather sets of words (synonyms, singular/plurals, etc.) that all contribute to the occurrence count stored in one component of the vector. In standard information retrieval this is dealt with via stemming (conflation) and a thesaurus.

The dimension m of the vector (that is, the number of relevant words associated to a context) is not theoretically limited to be particularly small, but one must keep in mind that in order to apply the method described over a significant number of pages is it reasonable to consider $m \leqslant 100$.

According to results of recent works [23,24], that have demonstrated that the use of the structures of HTML document improves the effectiveness of retrieving HTML documents, several extensions to the simple occurrence count have also been considered in the project partially described in this paper. Here we mention only that a method that identifies also the portion of the web page where a word appears (Title, Text, Link, etc.) has been designed to produce more accurate vectors associated to a page and to a list of words.

We have used several methods to identify the list of m words that are associated with a context/profile, that can be divided into two main groups:

1. creation of the list from user knowledge;
2. extraction of the list from given sets of pages.

In the first case, the words are determined in a setup phase, when the search engine managers decide which are the contextes/profiles supported and what are the words that are representative of

that context/profile. This operation may be accomplished together with the users of the engine itself when a thematic engine is devoted to a specific environment (such as an association of companies, a large corporation, a community of users).

In the second case, the words are identified starting from an initial set of pages, that are used as training sample for a context/profile. From the given set of pages one extracts the words with larger total occurrence in the set, and select from the list significant words (skipping articles, verbs, etc.). In standard information retrieval this is usually dealt with a stop word list. Many on-line vocabulary capable of performing this operation automatically are presently available, e.g., Wordnet (http://www.cogsci.princeton.edu/wn/) and Roget's (http://www.bartleby.com/62/). This approach is particularly significant in the case of the user profiles: in fact, one may consider as a training sample for a profile the pages that have been visited more recently by the user(s) that belong to that profile. This would allow the words associated with the profile to evolve with the behavior of the users in a smooth way.

The latter approach may be used to determine the relevant keywords for a set of pages on-line: when a query returns a set of results, the m most frequent valid words can be identified and then used to build the vectors. That operation can be done once all valid words appearing in each page have been properly stored in a database when the page is first acquired in the engine.

4. Page clustering

Document clustering in information retrieval is usually dealt with agglomerative hierarchical clustering algorithms or k-means algorithm.

Hierarchical agglomerative clustering algorithms [25] create a hierarchy of clusters, that may be viewed as a tree where each node is a cluster of objects and its children form a complete partition of that cluster. At start, a cluster for each object is created; then the method iterates through the cluster set by selecting the closest pair of clusters and merging them together forming a new cluster that replaces the two old clusters in a single one. The procedure is executed until all objects are contained within a unique cluster; then, clusters are selected by setting a threshold level on the inter-clusters distance. These methods attempt to minimize the sum of squares (SS) of the two clusters that can be formed at each step.

The k-means algorithm [26] starts by defining k cluster centroids or *seeds* and assigns each object to the cluster whose centroid is closest. Then, the algorithm is iterated, recomputing the cluster centroids and assigning the objects until a stable clustering is reached. The final solution depends on the choice of the initial centroids and on the parameter k, that determines the number of final clusters and has to be defined a priori.

In this study we have initially considered six different hierarchical agglomerative clustering methods that differ in the way they compute distances between clusters (single linkage, complete linkage, group average, weighted pair-group average, centroid, weighted centroid, and Ward's method) and a standard implementation of k-means. While agglomerative hierarchical clustering algorithms are very slow when applied to large document databases [27] (single link and group average methods take $O(|P|^2)$ time, complete link method take $O(|P|^3)$ time), k-means is much faster (its execution time is $O(k \cdot |P|)$).

For the latter, we have tested three different techniques for the determination of initial centroids: the first one selects the initial centroids by randomly choosing k documents from the data set, in the second the initial centroids are determined by randomly generating the elements of each centroid-vector, while in the third one the first centroid is randomly selected and a new centroid is iteratively selected from the other objects maximizing the total distance from the centroids already selected.

Measuring clustering effectiveness and comparing performance of different algorithms is a complex task, and there are no completely satisfactory methods for comparing the quality of the results of a clustering algorithm. In the experiments presented in this work we have used the Calinski–Harabasz (C–H) *pseudo-F* statistic as a measure of clustering quality; the higher the index value the better the cluster quality. For a given clustering, the mathematical expression of the *pseudo-F* statistic is

$$C\text{–}H = \frac{R^2}{(k-1)} \Big/ \frac{(1-R^2)}{(n-k)},$$

where

- $R^2 = (SST - SSE)/SST$;
- SST is the sum of the squared distances of each object from the overall centroid, i.e., $\sum_{j=1}^{k} \sum_{i=1}^{n_j} (x_{ij} - \bar{x})^2$ where x_{ij} is the value of the ith object in cluster j, n_j is the number of elements in cluster j, and $\bar{x}$ is the value of the overall centroid;
- SSE is the sum of the squared distances of each object from the centroid of its own group, i.e., $\sum_{j=1}^{k} \sum_{i=1}^{n_j} (x_{ij} - \bar{x}_j)^2$, where x_{ij} is defined as in SST and $\bar{x}_j$ is the value of the centroid of cluster j.

As a direct measure of the quality of clustering, this statistic is often used to identify the number of clusters in hierarchical methods. In a simulation study involving 30 different halting criterions for cluster analysis, Milligan and Cooper [28] found that the *pseudo-F* statistics was the one recovering the correct number of groups most often.

Table 1 summarizes some of the experiments conducted on data sets of real web documents. We have used 45 words for page vectorization; the results presented refer to two data sets containing 500 web pages and one containing 2633 pages (DS500_1 and DS500_2, DS2633 in the table). The number of clusters is fixed to 10. The last three rows of the table refer to the different methods to select the initial seeds for k-means described above. The results indicate that the quality of the k-means clustering with the first type of seed (k-Means Bar.1 in the table), as measured by the *pseudo-F* statistic, is always comparable with the one of the other slower methods. Considering the nature of our application, we have adopted k-Means Bar.1 due to its lower computational complexity.

5. The role of genetic algorithms

Our aim is to select a small subset P' of the original set of pages P for which the sum of the scores is large but also where the similarity amongst the selected page is restrained.

We select such a subset using a genetic algorithm (GA). Several reasons motivate this choice. First, the use of metaheuristic techniques is well established in optimization problems where the

Table 1
Comparison of clustering algorithms. The number of clusters is fixed to 10

Method	DS500_1 Pseudo-F	Time	DS500_2 Pseudo-F	Time	DS2633 Pseudo-F	Time
Single link	78.15	3	60.24	3	170.31	832
Complete link	103.72	4	78.70	4	230.26	759
Group average	80.64	4	62.36	3	208.82	676
Weighted average	71.23	3	61.72	4	307.15	1448
Unweighted centroid	62.48	4	62.36	4	165.63	1012
Weighted centroid	78.72	4	60.24	3	267.12	533
Ward's method	95.10	4	62.72	4	388.28	1286
k-Means bar.1	74.05	1	66.58	1	337.48	18
k-Means bar.2	99.12	1	65.73	1	293.28	20
k-Means bar.3	71.16	1	71.51	2	200.98	21

Times are expressed in seconds of CPU, PC with 1 MHz processor.

objective function and the constraints do not have a simple mathematical formulation. Second, we have to determine a good solution in a small computing time, where the dimension of the problem may be significantly large. Third, the structure of our problem is straightforward representable by the data structure commonly used by a GA.

GA (see e.g. [29–32]) are local search algorithms that start from an initial collection of strings (a population) representing possible solutions of the problem. Each string of the population is called a chromosome, and has an associated value called the *fitness function* (*ff*) that contributes in the generation of new populations by means of genetic operators. Every position in a chromosome is called a *gene* and its value is called the *allelic value*. This value may vary on an assigned allelic alphabet; often the allelic alphabet is $\{0, 1\}$. At each generation, the algorithm uses the fitness function values to evaluate the survival capacity of each string i of the population using simple operators in order to create a new set of artificial creatures (a new population) which try to improve on the current *ff* values by using pieces of the old ones. We summarize below the operators adopted.

- *Reproduction.* Individual strings are copied according to their objective function (*ff*) values. This represents a measure of the utility or goodness related to what we want to maximize. Copying strings according to their fitness function values means that strings with a high value have a higher probability of contribution to one or more offspring in the next generation.
- *Crossover.* The second operator denoted (simple) crossover works as follows: the members reproduced in the new mating pool are mated randomly and afterward each pair of strings, say a and b, undergoes a cross change. In order to do this, an integer position (cut point) is selected uniformly at random along the string. Splitting the two selected strings at this cut point produces a left part and a right part. An offspring can then be produced for example by joining the left part of a's encoding to the left part of b's encoding, or the right part of a's encoding to the right part of b's. Fig. 2 shows a simple crossover operation.
- *Mutation.* The mutation operator plays a secondary role with respect to reproduction and crossover operators. Nevertheless mutation is needed to prevent an irrecoverable loss of potentially useful

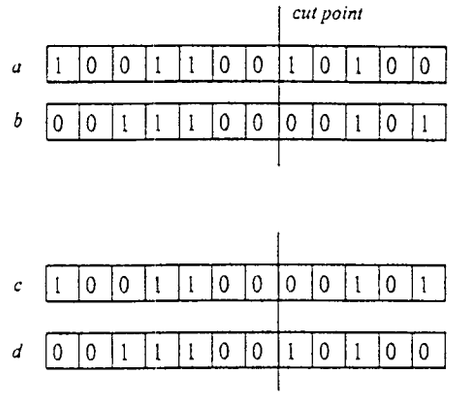


Fig. 2. A simple crossover operation: c and d are the two new strings produced by a and b. In the example the allelic alphabet is $\{0, 1\}$.

information which occasionally reproduction and crossover can cause. This operator is an occasional random alteration, of the allelic value of a chromosome that occurs with small probability.

Here follows the outline of the basic genetic algorithm:

1. Generate a-random population of n chromosomes (suitable solutions for the problem).
2. Evaluate the fitness ff of each chromosome in the population.
3. Create a new population by repeating the following steps until the new population is complete:
   - Select two parent chromosomes from a population according to their fitness (the better fitness, the bigger chance to be selected).
   - With a crossover probability cross over the parents to form new offspring (children). If no crossover was performed, offspring is the exact copy of parents.
   - With a mutation probability mutate new offspring at each position in chromosome.
   - Place new offspring in the new population.
4. Use new generated population for a further run of the algorithm.
5. If the end condition is satisfied, stop, and return the best solution in the current population.
6. Go to step 2.

In the following we discuss the algorithm engineering.

- *The population.* Starting from the clusters obtained, we define the chromosomes of the initial population as subsets of pages of bounded cardinality (in the GA terminology, a page is a *gene*). The genetic algorithm works on the initial population ending up with a representative subset of pages to present to the user. The idea is to start the genetic evolution with a population that is already smaller than the initial set of pages P. Each chromosome is created by picking a page from each cluster, starting with the ones with higher score. Thus, the first chromosome created will contain the pages with the highest score in each cluster, the second chromosome will contain the second best, and so on. If the cardinality of a cluster is smaller than the number of chromosomes to be created, then that cluster will not be represented in each chromosome, while other clusters with higher cardinality may have more that one page representing them in some chromosome.

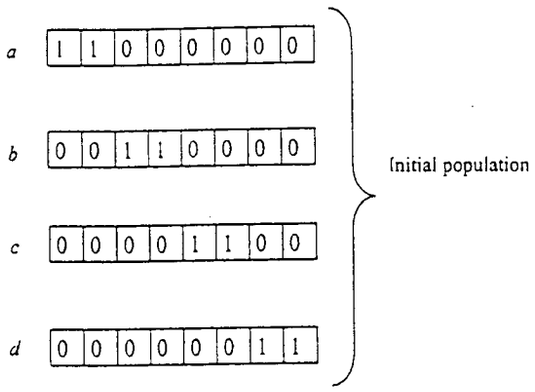


Fig. 3. An example of initial population formed by four chromosomes from a set of eight pages.

We indicate with dc the number of pages included in each chromosome in the initial population, and with nc the number of chromosomes. The population will thus contain $np = dc \cdot nc$ pages. Independent of which cluster the chromosome belongs, the allelic alphabet of its genes is formed by two binary elements, i.e., 0 and 1. Each gene represents a page and the values 0 or 1 assumed by the gene define whether that page is in the pool of pages represented by that chromosome or not. In other words we can say that a chromosome $l = 1, \ldots, ng$ is a vector whose entries are 0 or 1, according to the presence or the absence of that page in that chromosome.

Here follows a simple example. We have a set of eight pages, i.e., $P = \{1, \ldots, 8\}$, and the page clustering process with $k = 2$ (i.e., with two clusters) gives the following result: pages 1, 3, 5 and 7 are in the first cluster, while pages 2, 4, 6 and 8 are in the second cluster. Assuming, without loss of generality, that $score(i) \geqslant score(j)$ for all $i, j \in P$ such that $i < j$, an initial population of four chromosomes can be constructed as depicted in Fig. 3.

- *The fitness function.* The fitness function computed for each chromosome is expressed as a positive value that is higher for "better" chromosome, and is thus to be maximized. It is composed of three terms.

The first is the sum of the score of the pages in chromosome C, i.e.,

$$t_1(C) = \sum_{p_i \in C} score(p_i),$$

where $score(p_i)$ is the original score given to page p_i as described at the beginning of Section 2. This term considers the positive effect of having as many pages with high score as possible in a chromosome, but would also reward chromosomes with many pages regardless of their score (a chromosome with many pages with low score could produce a higher fitness of another with few good pages).

This drawback is balanced with the second term of the fitness function, that penalizes the distance of the dimension of a chromosome from an *ideal* dimension. Let ID be such ideal dimension; the ratio

$$t_2(C) = \frac{np}{abs(|C| - ID) + 1}$$

constitutes the second term of the fitness function: it reaches its maximum np when the dimension of C is exactly equal to the ideal dimension ID, and rapidly decreases when the number of pages contained in chromosome C is smaller or greater than ID.

The chromosomes that are present in the initial population are characterized by the highest possible variability as far as the clusters to which the pages belong are concerned. Although, the evolution of the population may alter this characteristic, creating chromosomes with high fitness where the pages belong to the same cluster and are very similar to each other. Moreover, the fact that pages belonging to different clusters are different in the vectorized space may not be guaranteed, as it depends both on the nature of the data and on the quality of the initial clustering process. For this reason, we introduce in the fitness function a third term, that measures directly the overall dissimilarity of the pages in the chromosome. Let $D(p_i, p_j)$ be the Euclidean distance of the vectors representing pages p_i and p_j. Then,

$$t_3(C) = \sum_{p_i, p_j \in C, p_i \neq p_j} D(p_i, p_j)$$

is the sum of the distances between the pairs of pages in chromosome C and measures the total variability expressed by C.

The final form of the fitness function for chromosome C is then

$$ff(C) = \alpha \cdot t_1(C) + \beta \cdot t_2(C) + \gamma \cdot t_3(C)$$

where α, β and γ are parameters that depend on the magnitude of the initial score and of the vectors that represent the pages. In particular, α, β and γ are chosen so as the contributions given by $t_1(C)$, $t_2(C)$ and $t_3(C)$ are balanced. Additionally, they may be tuned to express the relevance attributed to the different aspects represented by the three terms.

The goal of the GA is to find, by means of the genetic operators, a chromosome C^* such that $ff(C^*) = \max_{C=1,\ldots,nc} ff(C)$.

6. Experimental results

We have run a number of experiments with the method described above. The code implementing the clustering and the GA was developed in the standard C language under a Windows NT platform. The complete solution process for the instances described below took at most few seconds using a commercial PC with 1 GHz processor and 128 MHz RAM. The experiments are presented below in two subsections, the first devoted to experiments with randomly generated data, and the second dealing with a real application of the proposed method in an experimental thematic search engine.

6.1. Tests on randomly generated data

To verify the behavior of the method, we have generated at random a set of instances of different dimensions in the following way. Given the dimension of set P, say np, and the number m of keywords used for vectorization, we determine the matrix of occurrences with np rows and m columns in two steps. First, we fix a value ranging from 0.1 to 0.2 that expresses the density of the matrix (md). Then, we generate the entries of the matrix at random in the interval (1 – 50)

according to the uniform distribution, respecting the constraint on the density. The score of each page is itself determined at random and between 0 and 100. The main parameters of the algorithm have been tuned on a few test cases and then were fixed for all the experiments presented, in order to focus the variability of the results on the performance of the method for a given set of parameters rather than on the effect of varying the parameters. The ideal size of the subset of pages selected was set to 10 as to represent a typical output page of a standard search engine. As far as the values of the parameters are concerned, we have that:

- the number of clusters was set to 5 and the dimension of the initial chromosomes is 200;
- the total number of chromosomes was set to 40, resulting in total of $200 = 5 \cdot 40$ pages composing the initial population;
- α was set to 1;
- β was set to 50;
- γ was set to 1.

The first set of experiments was designed to test the behavior of the GA in determining a reduced set of pages with high fitness. The genetic algorithm that we have designed composes new chromosomes with high fitness in few iterations. Then, the highest fitness is seldom increased in the following iterations, despite the crossover and the mutation operations. This is shown in Table 2, where we report the increase in the fitness function value determined by the algorithm and the iterations at which the best chromosome has been created.

The total number of iterations was bounded to 1000. We have considered instances with 500, 1000, and 10 000 pages. In Table 2 the first column contains the name of the experiment, that reports the size of the initial set of pages and an index associated with each random replication of a data set with same dimensions.

From the results of Table 2 we see that a significant increase in the value of the fitness function is always determined (column six), and, moreover, the number of iterations needed to identify the best solution is small and often very small. Thus, we have reasonably resorted to halting the algorithm once the best fitness value has remained the same for the last 50 iterations.

In the second set of experiments, we have focused on the evaluation of the results obtained. For each experiment, we have compared the subset of pages determined by our method (referred to, in the following, as *genetic subset*) with the subset of the same dimension containing the best pages according to the initial score (referred to, in the following, as *standard subset*). Assume that the method has determined a subset of h pages to present to the user. We question what we lose and what we gain in presenting this set of pages instead of the best h pages according to the score, as in a standard query result. To assess the difference between the two sets, we consider and compare the following measures:

(a) the sum of the score of the pages in the subset;
(b) number of clusters represented in the subset (a cluster is represented in a subset if at least one page belonging to that cluster is in the subset);
(c) the average Euclidean distance amongst all pairs of pages in the subset;
(d) the average Hamming distance amongst all pairs of pages in the subset (the Hamming distance between two pages is computed as the sum of keywords for which some occurrences are present in one page and not present in the other).

Table 2
Behavior of the genetic algorithm for different randomly generated data sets

Experiment	Number of keywords	Matrix density	Best fitness at iteration 0	Best fitness overall	Fitness increase ratio	Best iteration
EXP500_1	20	0.10	1540.10	6115.79	2.97	368
EXP500_2	20	0.15	1549.64	6122.12	2.95	217
EXP500_3	20	0.20	1564.93	6140.98	2.92	133
EXP500_4	30	0.10	1564.81	6130.30	2.92	368
EXP500_5	30	0.50	1560.14	6137.34	2.93	48
EXP500_6	30	0.20	1568.80	6143.90	2.92	48
EXP500_7	50	0.10	1574.39	6141.66	2.90	48
EXP500_8	50	0.15	1567.95	6132.94	2.91	528
EXP500_9	50	0.20	1587.64	6171.67	2.89	49
EXP1000_1	20	0.10	1547.19	6127.88	2.96	205
EXP1000_2	20	0.15	1549.15	6129.54	2.96	165
EXP1000_3	20	0.20	1561.25	6148.16	2.94	46
EXP1000_4	30	0.10	1564.10	6148.63	2.93	56
EXP1000_5	30	0.50	1567.21	6154.22	2.93	194
EXP1000_6	30	0.20	1572.41	6164.46	2.92	51
EXP1000_7	50	0.10	1574.04	6146.77	2.90	99
EXP1000_8	50	0.15	1581.33	6172.41	2.90	48
EXP1000_9	50	0.20	1595.10	6191.06	2.88	79
EXP10000_1	20	0.10	1550.91	6149.68	2.86	988
EXP10000_2	20	0.15	1550.91	6149.68	2.96	50
EXP10000_3	20	0.20	1550.91	6149.68	2.96	50
EXP10000_4	30	0.10	1571.79	6170.93	2.93	88
EXP10000_5	30	0.50	1575.34	6185.65	2.93	349
EXP10000_6	30	0.20	1581.10	6173.52	2.90	545
EXP10000_7	50	0.10	1581.06	6182.25	2.91	198
EXP10000_8	50	0.15	1603.56	6205.47	2.87	227
EXP10000_9	50	0.20	1610.41	6218.97	2.86	988

While (a) measures the quality of the subset according to page score, term (b) aims to evaluate how the final subset takes into account the structure of the initial set of pages (represented by the clusters). Terms (c) and (d) are both indirect measures of the dissimilarity among the pages in the subset. While the average Euclidean distance is itself weighted in the fitness function used in the GA, we have introduced the additional measure of dissimilarity based on the Hamming distance to make the evaluation of the quality of the subset presented more stable. In fact, although the two measures are likely to be correlated, the second one has been often used as a consistent indicator of similarity between documents.

In the results of Table 3 we have fixed the number of keywords at 30 and the density of the matrix of occurrences to 0.15, and have instead replicated different randomly generated instances for 500, 1000, and 10000 pages. The last three columns of the table report the percentage of the variation

Table 3
Evaluation of results on randomly generated instances

Experiment	Dimension of best chromosome	Cluster represented Standard	Genetic	Total score % var.	Avg. Hamming distance % var.	Avg. Euclidean distance % var.
EXP500_1	11	3	5	−1.7	22.5	1.3
EXP500_2	11	5	5	−1.5	21.6	0.3
EXP500_3	11	3	5	−5.4	13.1	12.7
EXP500_4	11	4	5	−1.7	4.7	0.0
EXP500_5	11	4	5	−3.1	20.3	2.7
EXP500_6	11	5	5	−1.5	21.6	27.0
EXP500_7	11	5	5	−2.2	6.1	5.4
EXP500_8	11	5	5	−2.4	10.3	1.3
EXP500_9	11	5	5	−2.1	35.7	17.9
EXP500_10	11	2	4	−3.9	21.6	8.4
Average_500	11	4.1	4.9	−2.5	17.7	7.7
EXP1000_1	11	3	5	−1.5	23.7	16.1
EXP1000_2	11	4	5	−0.2	1.6	5.4
EXP1000_3	11	3	5	−0.8	5.3	5.7
EXP1000_4	11	4	5	−0.6	6.8	15.0
EXP1000_5	11	5	5	−1.6	6.8	17.8
EXP1000_6	11	3	5	−1.5	0.0	14.8
EXP1000_7	11	4	5	−2.2	14.5	24.7
EXP1000_8	11	5	5	−0.9	28.0	22.5
EXP1000_9	11	5	5	−0.8	20.2	8.3
EXP1000_10	11	4	5	−0.6	11.4	16.9
Average_1000	11	4	5	−1.1	11.8	14.7
EXP10000_1	11	4	5	−0.1	16.5	16.9
EXP10000_2	11	4	5	0.5	14.7	15.7
EXP10000_3	11	4	5	0.0	5.7	7.2
EXP10000_4	11	3	4	−0.6	0.2	19.0
EXP10000_5	11	4	5	−0.1	31.7	16.9
EXP10000_6	11	5	5	−0.1	23.7	24.3
EXP10000_7	11	4	5	−0.5	44.1	34.6
EXP10000_8	11	3	5	−0.4	36.3	21.1
EXP10000_9	11	4	3	−0.3	69.2	38.0
EXP10000_10	11	4	5	−0.4	32.3	42.0
Average_10,000	11	3.9	4.7	−0.2	27.4	23.6

obtained with the *genetic subset* with respect to the *standard subset*. A positive value means that genetic is better than standard, while a negative one indicates the opposite.

It is evident from the results of Table 3 how the solutions identified by our method provide a significant gain in page dissimilarity (measured both by Euclidean and Hamming average distance, columns five and six of the table) while the loss in total score with respect to the simple ordering is negligible (column four in the table).

Table 4
Evaluation of results on randomly generated instances with page replications

Experiment	Dimension of best chromosome	Number of duplicates	Total score % var.	Avg. Hamming distance % var.	Avg. Euclidean distance % var.
EXP500_1	11	0	−2.2	6.1	5.4
EXP500_2	11	1	−1.9	21.4	8.0
EXP500_3	11	2	−2.1	28.3	21.3
EXP500_4	11	3	−1.8	16.8	8.3
EXP1000_1	11	0	−1.0	7.9	4.3
EXP1000_2	11	1	−0.9	10.25	9.7
EXP1000_3	11	2	−0.8	12.3	8.1
EXP1000_4	11	3	−1.1	21.8	21.7
EXP10000_1	11	0	−0.4	30.2	33.3
EXP10000_2	11	1	−0.4	34.2	34.2
EXP10000_3	11	2	−0.4	58.9	52.8
EXP10000_4	11	3	−0.4	58.9	56.6

A third set of experiments (see Table 4) has been designed to evaluate the performance of the algorithm when duplicated pages are present. For the same setting of the parameters as above, for different dimensions of the initial set of pages, we have considered the same instance and have artificially duplicated the page with best score once, twice, and three times. The objective of the experiments is to assess the ability of the method of excluding duplications from the final results. The figures associated to the measures of dissimilarity between pages in Table 3 are promising: in fact, they show that, when duplications are present, the total variability of the subsets selected by our algorithm increases with respect to the variability of the subset determined with the simple ordering. This effect, being the data set exactly the same except for the duplicated pages, is to be attributed to the fact the duplications appear in the classic ordering while are excluded by the GA in our solutions. This behavior is more evident when the number of duplicated pages is increased.

6.2. *Tests on real data*

The method described in this paper has been developed in the framework of a project funded by the Italian Ministry of Industry; within the same project a complete thematic search engine has been realized and is now being integrated with the new search method.

Using this tool, a thematic database of web pages on "Infancy & Childhood" was collected through spidering Italian web sites in the first months of 2003. Out of this database we have extracted 3710 pages, that were the result of a specific search on the database of the engine.

The pages were then vectorized according to 57 macrokeywords, where each such keyword was effectively the collection of the occurrences of root words and strong synonyms; this determined a vector of 57 integers for each page that described that page for the scopes of our application.

Table 5
Results obtained on a database of 3710 web pages

Experiment	Dimension of best chromosome	Cluster represented Standard	Genetic	Total score % var.	Avg. Hamming distance % var.	Avg. Euclidean distance % var.
User score 1	10	1	5	−7.77	35.38	747.61
User score 2	11	2	5	−9.09	16.97	433.92
User score 3	10	1	5	0.00	23.58	980.13
User score 4	11	3	5	−10.82	0.10	266.01
User score 5	10	1	5	−5.83	32.19	646.29

The 3710 pages were then sorted according to five different types of user score, corresponding to different preference criteria expressed by test users in terms of the macro keywords. The algorithm was set up selecting a total of 200 chromosomes each with 1000 genes and 5 pages assigned resulting in an initial "pool" of the best 1000 web pages according to the specific user score adopted. Then, the algorithm was run to find the best chromosome, i.e., the best subset of pages, according to the procedures described in the previous sections.

The results of the experiments are presented in Table 5. The columns of the table have the same format of the tables in Section 6.1, reporting the dimension of the best chromosome, the number of clusters represented, and the variations of

(i) the total user score,
(ii) the average Hamming distance, and
(iii) the average Euclidean distance,

between the pages in the selected chromosome and the pages obtained by the simple ordering of the user score.

Also in this case the behavior of the algorithm is interesting: small reductions in the total score of the subset are balanced by high increases in the clusters representativeness and internal dissimilarity in the selected pages.

7. Conclusions

The experimental results presented in this paper show that the proposed method can be effective in the selection of small subsets of pages of good quality, where quality is not considered as a simple sum of the quality of each page but as a global characteristic of the subset. The implementation of the GA and of the clustering algorithm have allowed us to obtain convergence to solutions in reasonably short computational times on a standard personal computer (a few seconds).

One may question whether the described method can be run on-line in a search engine as the standard execution of a user's query. We believe that with proper tuning on the parameters and a proper engineering of the algorithms, the computational effort can be dealt with and that only few seconds may be added to the overall search process. Future work will cover the extension of the

page vectorization technique and the definition and test of automatic procedures for parameter tuning in the genetic algorithm.

References

[1] Glover EJ, Lawrence S, Gordon MD, Birmingham WP, Giles CL. Recommending web documents based on user preferences. SIGIR 99 Workshop on Recommender Systems, Berkeley, CA.

[2] Glover EJ, Lawrence S, Gordon MD, Birmingham WP, Giles CL. Web search—your way. Communications of the ACM 2001;44(12):97–102.

[3] Willet P. Recent trends in hierarchical document clustering: a critical review. Information Processing & Management 1988;24(5):577–97.

[4] Cutting DR, Karger DR, Pedersen JO, Tukey JW. Scatter/Gather: a cluster-based approach to browsing large document collections. In: Proceedings of the 15th Annual International ACM/SIGIR Conference, Copenhagen, 1992.

[5] Leuski A. Evaluating document clustering for interactive information retrieval. In: Proceedings of the 2001 ACM CIKM International Conference on Information and Knowledge Management, Atlanta, Georgia, USA, November 2001. p. 33–44.

[6] Zamir O, Etzioni O, Madani O, Karp RM. Fast and intuitive clustering of web documents. KDD '97, Newport Beach, CA, 1997. p. 287–90.

[7] Zamir O, Etzioni O. Grouper: a dynamic clustering interface to web search results. WWW8/Computer Networks 1999;31(11–16):1361–74.

[8] Koller D, Sahami M. Hierarchically classifying documents using very few words. In: Proceedings of the 14th International Conference on Machine Learning, Nashville, Tennessee, July 1997. p. 170–8.

[9] Aggarwal CC, Gates SC, Yu PS. On the merits of building categorization system by supervised clustering. In: Proceedings of the Fifth ACM SIGKDD International Conference on Information and Knowledge Management, Bethesda, Maryland, USA, November 1998. p. 3–7.

[10] Van Rijsbergen CJ. Information retrieval, 2nd ed. London: Butterworths, 1979.

[11] Hearst MA, Pederson JO. Reexamining the cluster hypothesis: scatter/gather on retrieval results. In: Proceedings of ACM SIGIR '96, Zurich, August, 1996.

[12] Gordon M. Probabilistic and genetic algorithms for document retrieval. Communications of the ACM 1988;31(10): 1208–8.

[13] Gordon, M. User-based document clustering by redescribing subject descriptions with a genetic algorithm. Journal of the American Society for Information Science 1991;42(5):311–22.

[14] Yang JJ, Korfhage RR. Query optimization in information retrieval using genetic algorithms. In: Proceedings of the Fifth ICGA, Urbana, Champaign, IL, USA, 1993. p. 603–13.

[15] Chen H. Machine learning for information retrieval: neural networks, symbolic learning, and genetic algorithms. Journal of the American Society for Information Science 1995;46(3):194–216.

[16] Kraft DH, Petry FE, Buckles BP, Sadavisan T. Genetic algorithms for query optimization in information retrieval: relevance feedback. In: Sanchez E, Zadeh LA, Shibata T, editors. Genetic algorithms and fuzzy logic systems, soft computing perspectives. Singapore: World Scientific, 1997. p. 155–73.

[17] Petry F, Buckles B, Prabhu D, Kraft D. Fuzzy information retrieval using genetic algorithms and relevance feedback. In: Proceedings of the ASIS Annual Meeting, Columbus, OH, 1993. p. 122–5.

[18] Sanchez E, Pierre Ph. Fuzzy logic and genetic algorithms in information retrieval. In: Proceedings of the Third International Conference on Fuzzy Logic, Neural Net and Soft Computing, Iizuka, Japan, 1994. p. 29–35.

[19] Yang JJ, Korfhage RR, Rasmussen E. Query improvement in information retrieval using genetic algorithms: a report on the experiments of the TREC project. In: Proceedings of the First Text REtrieval Conference (TREC-1), Gaithersburg, MD, 1993. p. 31–58.

[20] Boughanem M, Chrisment C, Tamine L. Genetic approach to query space exploration. Information Retrieval 1999;1:175–92.

[21] Horng J-T, Yeh C-C. Applying genetic algorithms to query optimization in document retrieval. Information Processing and Management 2000;36:737–59.

[22] Zacharis ZN, Panayiotopoulos T. Web search using a genetic algorithm. IEEE Internet Computing March/April 2001;5(2):18–26.
[23] Cutler M, Deng H, Maniccam SS, Meng W. A new study on using HTML structures to improve retrieval. In: Proceedings of the 11th IEEE International Conference on Tools with Artificial Intelligence, Chicago, Illinois, 1999.
[24] Cutler M, Shih Y, Meng W. Using the structure of HTML documents to improve retrieval. USENIX Symposium on Internet Technologies and Systems (NSITS'97), Monterey, California, 1997. p. 241–51.
[25] Jain AK, Murty MN, Flynn PJ. Data Clustering: a review. ACM Computing Surveys 1999;31(3):264–323.
[26] Dubes RC, Jain AK. Algorithms for clustering data. Englewood Cliffs, NJ: Prentice-Hall, 1988.
[27] Zamir O, Etzioni O. Web document clustering: a feasibility demonstration. In: Proceedings of 19th International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR 98), Melbourne, Australia, 1998. p. 46–54.
[28] Milligan GW, Cooper M. An examination of procedures for determining the number of clusters in a data set. Psychometrika 50(2):159–79.
[29] Chambers L. Practical handbook of genetic algorithms: applications. Boca Ratoon, Florida: CRC Press, 1995.
[30] Goldberg ED. Genetic algorithms in search optimization & machine learning. Reading, MA: Addison-Wesley Publishing Company, 1999.
[31] Holland JH. Adaptation in natural and artificial systems. Ann Arbor, MI: The University of Michigan Press, 1975.
[32] Michalewicz Z. Genetic algorithms + data structures = evolution programs. Berlin, Heidelberg: Springer, 1992.

Modelling the Limber Thesaurus in RDF

Brian Matthews
CLRC-RAL
Ken Miller
UKDA

1. INTRODUCTION AND PURPOSE

In this document we consider the model of the information held in the ELSST thes being developed as part of the LIMBER project.

We give a model for the in terms of the general RDF model for defining metadat propose a serialisation of that model in RDF, documenting the constraints on the n giving an RDF schema and some examples RDF instances.

This model follows the structure of thesaurus content as defined in the ISO standard and 5964. It is modified from earlier work by Phil Cross and Dan Brickley (ILRT B and Traugott Koch (NETLAB,Lund, Sweden).

This model will form the basis of defining an API onto thesaurus information, to be by the Thesaurus Management System and Indexing Tool to be developed v LIMBER.

2. INTRODUCTION TO RDF

We give a brief overview of the W3C Resource Description Framework.

3. MODELLING THESAURUS DATA

The thesaurus contains a set of *concepts*. Each concept is seen as a single unit of me in the thesaurus. Each concept has a unique *classification code*.

Each concept *is indicated by* a set of *terms* defining words and phrases which represe concept. Each concept can have more than one term; each term is in one or more co (The restriction is neither useful nor necessary. Homonyms are frequent, and use thesaurus is severely restricted by that condition). Each term has a *language* and a Languages are given by a reference to a standard language code given in ISO 639, ar value by a *literal string*. Certain terms are *preferred terms* and represent the canonical for that concept; there should be at most one preferred term for each language.

Deleted: and only one

A concept can have a set of *scope notes*. A scope note has a value, being a literal str language, giving the language of the notes, and a *scope note type*, three of whic defined: *Regular, Translation, Hierarchy*. The Getty Information Institute uses also: " note" and "history note". Only top concepts (see below) can have hierarchy scope not

leted: a tree structure

Concepts are arranged in *hierarchies*. At the top of the hierarchy is a *top concept*. Belo top concept, concepts are arranged in, in a DAG. (restriction to tree structu counterproductive. Enough thesauri are of DAG type). . Each concept has a relation the unique top concept in its hierarchy.

The hierarchy should belong to a thesaurus. This could be denoted via the top-co The best is, to identify concepts as artifacts of a thesaurus, and terms as parts language. The concept-concept relation belong either to a thesaurus or to a corre between thesauri. In this proposal, ISO5964 is not reflected. It seems to be just assur implicit in the concept-concept relation. It is however important for practical manag to know who (which thesaurus editing authority) defines concepts and creates co concept relations. In a wider environment, an API for thesaurus access must be a identify a pool of concepts and terms from different thesauri.

Deleted: the

The parent concept in this tree is in the *broader concept relation* for that concept (mult a child concept is a *narrower concept relation*. If concept A is a narrower concept for co B then B must be a broader concept of A. All broader and narrower terms must sha same top concept. Better: All narrower terms must share the same top concept broader terms must be under at least one of the top concepts. This logic is valid and in TMS.

Additionally, any concept can have a *related concept* relation to any other conce concept A is a related concept of concept B, then B must be a related concept of A.

ISO5964: missing are: exact equivalence, inexact equivalence, one-to-many:AND, o many:OR, broader equivalence, narrower equivalence.

Two indexing languages can not be reduced to a common tree of concepts!

4. EXAMPLE

Figure 1 gives a diagrammatic representation of the Friend concept, with classifi code 620.

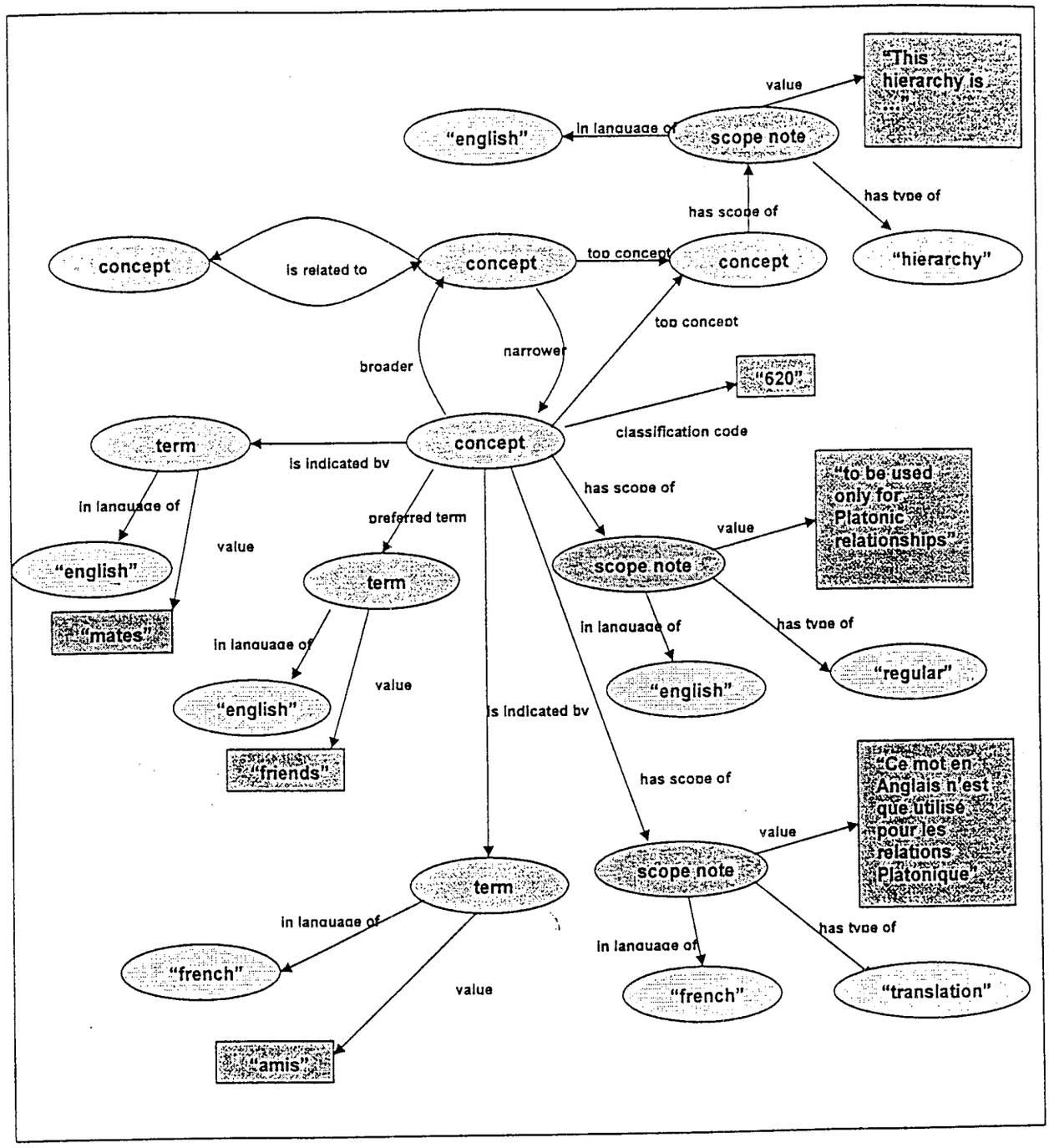


Figure 1: A RDF instance diagram for the multilingual *Friend* concept.

5. AN RDF SCHEMA FOR THESAURI

In this section, we give the RDF Schema for multilingual thesauri to express the r above. This follows the structure of thesaurus content as defined in the ISO standard: and 5964. It is modified from earlier work from Phil Cross from ILRT in Bristol.

Language identifiers are supplied by referring to the list of ISO639 identifiers, expres: RDF, with URI: http://www.itd.clrc.ac.uk/Projects/Limber/ISO639. This may char the project evolves - would be suitable to put in one common place.

Note that we do note use xml:lang as the language identifier here, but explicitly separate property.

```
<rdf:RDF xml:lang="en"
         xmlns:rdf="http://www.w3.org/1999/02/22-rdf-syntax-ns#"
         xmlns:rdfs="http://www.w3.org/2000/01/rdf-schema#">
<!--

     The RDF Schema for thesuari.

     Brian Matthews, Limber Project,

     Information Technology Department,
     CLRC Rutherford Appleton Laboratory

     25/10/00
-->
```

5.1 Classes

Only four fundamental classes of object are defined in the Thesaurus model.

There should be a class TopConcept. This carries the "hierarchy note". It belong: "thesaurus". In turn, all concepts and concept-concept relations below (except IS(ones) are regarded products of this thesaurus.

The top-level object is the *ThesaurusObject;* all other classes will be subclasses of this c This will allow us to put any common properties or constraints on all classes (if any).

```
<rdfs:Class rdf:ID="ThesaurusObject">
  <rdfs:comment>
     All classes will be an subclass of the top-level class Thesaurus
     concept.  This will allow us to put any common properties or
     constraints on all classes (if any).
  </rdfs:comment>
  <rdfs:subClassOf
       rdf:resource="http://www.w3.org/2000/01/rdf-schema#Resource" />
</rdfs:Class>
```

The *Concept* class describes as unique concept in the vocabulary, possibly defined v some scheme. such as a thesaurus or classification scheme. Instances can us

rdfs:isDefinedBy property with a vocabulary namespace as its value, to indicat vocabulary to which the concept belongs.

```
<rdfs:Class rdf:ID="Concept">
      <rdfs:comment>
         A unique concept defined within a vocabulary scheme, such as a
         thesaurus or classification scheme. Instances can use the
         rdfs:isDefinedBy property with a vocabulary namespace as its
         value, to indicate the vocabulary to which the concept
         belongs.
      </rdfs:comment>
      <rdfs:subClassOf rdf:resource="#ThesaurusObject"/>
</rdfs:Class>
```

The *Term* class captures a textual realisation of a concept, such as a word or phrase.

```
      <rdfs:Class rdf:ID="Term">
            <rdfs:comment>Instances of this class represent the written fori
        Concepts, capturing a word or phrase that expresses the concept.
        The string is given by the rdf:value of Term.
      </rdfs:comment>
            <rdfs:subClassOf rdf:resource="#ThesaurusObject"/>
      </rdfs:Class>
```

The *ScopeNote* class provides a comment on a concept, prescribing the boun appropriate usage of the concept.

```
      <rdfs:Class rdf:ID="ScopeNote">
            <rdfs:comment>Provides a comment on the concept, for disambigua
explanation etc.
          The string is given by the rdf:value of ScopeNote.
      </rdfs:comment>
            <rdfs:subClassOf rdf:resource="#ThesaurusObject"/>
      </rdfs:Class>
```

5.2 Properties

5.2.1 Properties of Concepts.

The *ClassificationCode* property provides a unique identifier for a concept.

```
      <rdfs:Property rdf:ID="ClassificationCode">
            <rdfs:comment>The unique identifier of a concept.</rdfs:comment:
            <rdfs:domain rdf:resource="#Concept"/>
            <rdfs:range
                  rdf:resource="http://www.w3.org/2000/01/rdf-schema#Litera:
      </rdfs:Property>
```

Unique identifiers may change, and many thesauri don't use them. If thes updates are to be transmitted, changes of identifiers and/or preferred term be foreseen!

The *hasScopeNote* property defines the relationship between concepts and scope notes

```
<rdfs:Property rdf:ID="hasScopeNote">
      <rdfs:comment>A scope note of a concept.</rdfs:comment>
      <rdfs:domain rdf:resource="#Concept"/>
      <rdfs:range rdf:resource="#ScopeNote"/>
</rdfs:Property>
```

The *isIndicatedBy* property defines the relationship between concepts and terms.

```
<rdfs:Property rdf:ID="isIndicatedBy">
      <rdfs:comment>A defining term for a concept.</rdfs:comment>
      <rdfs:domain rdf:resource="#Concept"/>
      <rdfs:range rdf:resource="#Term"/>
</rdfs:Property>
```

The *PreferredTerm* property is a subproperty of the *isIndicatedBy* property, defining term which is the reference term for the concept. The "usedfor" or "synonym" or preferred term" should be explicitly used as subproperty of isIndicatedBy to identi not "preferred term".

```
<rdfs:Property rdf:ID="PreferredTerm">
      <rdfs:comment>The preferred term for a concept w.r.t. a spe
language.</rdfs:comment>
      <rdfs:subPropertyOf rdf:resource="#isIndicatedBy"/>
</rdfs:Property>
```

The property *ConceptRelation* is a generalisation of all the semantic relationship bet properties. The Broader, Narrower, Top, and Related relationships are a subproperties.

```
<rdfs:Property rdf:ID="ConceptRelation">
      <rdfs:comment>A generalisation of all relationships be
concepts.</rdfs:comment>
      <rdfs:domain rdf:resource="#Concept"/>
      <rdfs:range rdf:resource="#Concept"/>
</rdfs:Property>

<rdfs:Property rdf:ID="BroaderConcept">
      <rdfs:comment>The broader concept relation.</rdfs:comment>
      <rdfs:subPropertyOf rdf:resource="#ConceptRelation"/>
</rdfs:Property>

<rdfs:Property rdf:ID="NarrowerConcept">
      <rdfs:comment>The narrower concept relation.</rdfs:comment>
      <rdfs:subPropertyOf rdf:resource="#ConceptRelation"/>
</rdfs:Property>

<rdfs:Property rdf:ID="TopOfHierarchy">
      <rdfs:comment>The top concept relation.</rdfs:comment>
      <rdfs:subPropertyOf rdf:resource="#ConceptRelation"/>
      <rdfs:range rdf:resource="#TopConcept"/>
</rdfs:Property>
```

```
      <rdfs:Property rdf:ID="isRelatedTo">
            <rdfs:comment>The related concept relation.</rdfs:comment>
            <rdfs:subPropertyOf rdf:resource="#ConceptRelation"/>
      </rdfs:Property>
```

5.2.2 *Properties of Terms*

The *inLanguageOf* property defines the specific language of the term. This is giv terms of a reference to a LanguageCode object in within the Language Codes.

```
      <rdfs:Property rdf:ID="inLanguageOf">
            <rdfs:comment>The language of a term.</rdfs:comment>
            <rdfs:domain rdf:resource="#Term"/>
            <rdfs:range
rdf:resource="http://www.itd.clrc.ac.uk/Projects/Limber/ISO639#LanguageCode
      </rdfs:Property>
```

5.2.3 *Properties of Scope Notes*

Defines the language the scope note is written in. Uses the RDF Language Code defin

```
      <rdfs:Property rdf:ID="inLanguageOf">
            <rdfs:comment>The language of a scope note.</rdfs:comment>
            <rdfs:domain rdf:resource="#ScopeNote"/>
            <rdfs:range
rdf:resource="http://www.itd.clrc.ac.uk/Projects/Limber/ISO639#LanguageCode
      </rdfs:Property>
```

Scope notes are of several different classifications or types.

```
      <rdfs:Property rdf:ID="hasTypeOf">
            <rdfs:comment>The type of a scope note.</rdfs:comment>
            <rdfs:domain rdf:resource="#ScopeNote"/>
            <rdfs:range rdf:resource="#ScopeNoteType"/>
      </rdfs:Property>
```

The RDF Schema also allows us to define specific values for the class ScopeNoteType

```
      <ScopeNoteType rdf:ID="General"/>
      <ScopeNoteType rdf:ID="Hierarchy"/>
      <ScopeNoteType rdf:ID="Translation"/>
      <ScopeNoteType rdf:ID="Editor note"/>
      <ScopeNoteType rdf:ID="History note"/>

</rdf:RDF>
```

5.3 Language Codes

ISO 639 language codes are given as RDF values in a separate file, as follows

```
<rdf:RDF xml:lang="en"
      xmlns:rdf="http://www.w3.org/1999/02/22-rdf-syntax-ns#"
      xmlns:rdfs="http://www.w3.org/2000/01/rdf-schema#"
>
```

```
<!--
Brian Matthews, Limber Project, 12/09/00

Information Technology Department,CLRC Rutherford Appleton Laboratory
-->

<!-- should be one of these for each language in ISO 639 Language Codes -->

<LanguageCode rdf:ID="de"/>
<LanguageCode rdf:ID="en"/>
<LanguageCode rdf:ID="es"/>
<LanguageCode rdf:ID="fr"/>
. . .
</rdf:RDF>
```

6. CONSTRAINTS ON THE MODEL

Additionally, we wish to constrain the thesaurus model with extra conditions o consistency of the thesaurus. These cannot be expressed directly in RDF, althougl could be added by extending the *ConstraintResource* class in the RDF Schema defir which we quote below:

> This specification defines a subclass of resources known as 'constraint reso (section 3.1). This is provided to allow for the addition of new ways of expr RDF constraints. Future extensions to the Resource Description Frame may introduce new resources that are instances of the rdfs:ConstraintRes class. It is necessary to anticipate RDF content which draws upon propert classes defined using constraints other than those available in this version of As yet unknown constraints may contribute to a more expressive framewo specifying RDF constraints.

Here, the constraints are given in using a formal model. T is a thesaurus including a concepts *Concept*, a set of scope note *ScopeNote* and a set of terms *Term*. Each prope the RDF model above becomes a function in the formal model, returning a *set* of cor in an obvious manner.

1. The immediately broader concept of a concept is unique. That is, every concept most one broader concept. This is not necessary. Rather, BT must be acyclic.

   $$\forall c_1, c_2 \in Concept \bullet \exists\, c \in Concept \bullet$$
   $$c_1 \in BroaderConcept\,(c) \wedge c_2 \in BroaderConcept\,(c) \Rightarrow (c_1 = c_2)$$

   Thus, without ambiguity, we can treat *BroaderConcept* as a partial function.

2. The top concept of a concept is unique. That is, every concept has at most or concept. Neither this offers any advantage. Each concept must have at leas TopConcept. No TopConcept has another TopConcept.

$\forall c_1, c_2 \in Concept \bullet \exists\, c \in Concept \bullet$

$c_1 \in TopOfHierarchy\,(c) \wedge c_2 \in TopOfHierarchy\,(c) \Rightarrow (\, c_1 = c_2\,)$

Thus, without ambiguity, we can treat *TopOfHierarchy* as a partial function.

3. Narrower and Broader concepts are "inverse". That is, a concept is a narrower co of its broader concept and the broader concept of all its narrower concepts.

$\forall c_1, c_2 \in Concept \bullet c_1 = BroaderConcept\,(c_2) \Leftrightarrow c_2 \in NarrowerConcept\,(c_1)$

4. There is a unique top concept for each hierarchy.

Define the set *Ancestors*: $\forall c,c' \in Concept \bullet BroaderConcept\,(c) \in Ancestors\,(c) \wedge$
$(\, c' \in Ancestors\,(c) \Rightarrow BroaderConcept\,(c') \in Ancestors\,(c)\,)$

then: $\forall c, c' \in Concept \bullet c' \in Ancestors\,(c) \Rightarrow TopOfHierarchy\,(c') = TopOfHierarchy\,($

5. Top concepts have no broader concepts.

$\forall c \in Concept \bullet (\exists\, c_1 \in Concept \bullet\ c\ = TopOfHierarchy\,(c_1)) \Rightarrow Ancestors(c) = \varnothing$

6. Language/preferred term pairs are unique to concepts. Right.

$\forall c \in Concept \bullet \forall\ t_1\ t_2 \in Term\ \bullet t_1 = PreferredTerm(\, c\,) \wedge t_2 = PreferredTerm(\, c\,)$
$\wedge\ LanguageCode\,(t_1) = LanguageCode\,(t_2\,) \Rightarrow t_1 = t_2$

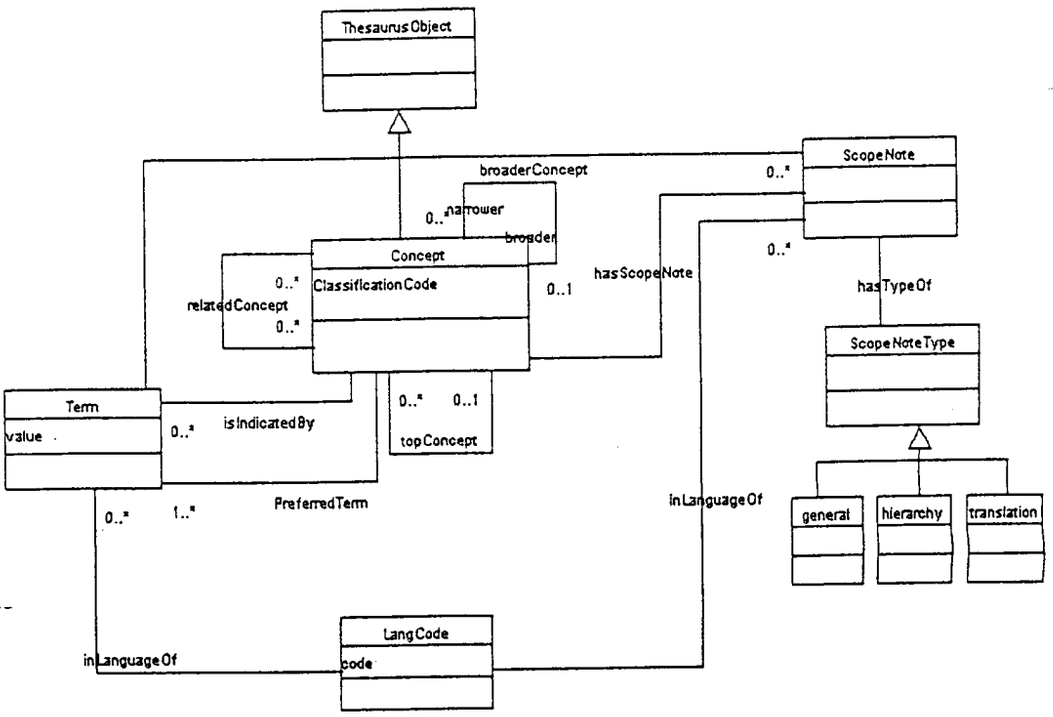
Thus for each concept, there can be at most one preferred term in any languag database terms, the pair preferred term - language code forms a unique key. This can replace unique codes.

7. Only top concepts have "hierarchy" scope notes. Note that they could be sin hierarchies. A TopConcept class makes life easier!

$\forall c \in Concept \bullet\ \forall n \in ScopeNote\ \bullet n \in ScopeNote\,(\, c\,) \wedge hasTypeOf(n) =$ "hierarchy"
$\Rightarrow Ancestors(c) = \varnothing$

7. AN EQUIVALENT UML MODEL

We give an equivalent UML class diagram in the figure below.



8. THE ABOVE EXAMPLE PRESENTED IN RDF

We represent the above example as an instance of the thesaurus RDF Schema.

This is a self-contained description of concepts, for transmission purposes and simp of presentation. To refer to concepts in for example Hasset, we would use something

```
<rdf:Description rdf:ID="http://www.data-archive.ac.uk/hasset/concepts/620"
    ...
    <rdfs:isDefinedBy
        rdf:resource="http://www.data-archive.ac.uk/hasset/concepts">
    ...
</rdf:Description>
```

```
<rdf:RDF xml:lang="en"
         xmlns:rdf="http://www.w3.org/1999/02/22-rdf-syntax-ns#"
         xmlns:rdfs="http://www.w3.org/2000/01/rdf-schema#">
<!--
```

```
Brian Matthews, Limber Project,

Information Technology Department,
CLRC Rutherford Appleton Laboratory

-->
      <rdf:Description rdf:ID="620">
            <rdf:type
rdf:resource="http://www.itd.clrc.ac.uk/Projects/Limber/thesaurus#Concept"/:
            <thes:ClassificationCode>620</thes:ClassificationCode>
            <thes:PreferredTerm>
                  <rdf:Description>
                        <rdf:type
rdf:resource="http://www.itd.clrc.ac.uk/Projects/Limber/thesaurus#Term"/>
                        <thes:inLanguageOf
rdf:resource="http://www.itd.clrc.ac.uk/Projects/Limber/ISO639#en"/>
                        <rdf:value>friends</rdf:value>
                  </rdf:Description>
            </thes:PreferredTerm>
            <thes:isIndicatedBy>
                  <rdf:Description>
                        <rdf:type
rdf:resource="http://www.itd.clrc.ac.uk/Projects/Limber/thesaurus#Term"/>
                        <thes:inLanguageOf
rdf:resource="http://www.itd.clrc.ac.uk/Projects/Limber/ISO639#en"/>
                        <rdf:value>mates</rdf:value>
                  </rdf:Description>
            </thes:isIndicatedBy>
            <thes:isIndicatedBy rdf:resource="#amis"/>
            <thes:hasScopeNote>
                  <rdf:Description>
                        <rdf:type
rdf:resource="http://www.itd.clrc.ac.uk/Projects/Limber/thesaurus#ScopeNote'
                        <thes:inLanguageOf
rdf:resource="http://www.itd.clrc.ac.uk/Projects/Limber/ISO639#en"/>
                        <thes:hasTypeOf
rdf:resource="http://www.itd.clrc.ac.uk/Projects/Limber/thesaurus#Regular"/:
                        <rdf:value>To    be    used    only    for    Pla
relationships.</rdf:value>
                  </rdf:Description>
            </thes:hasScopeNote>
            <thes:isIndicatedBy rdf:resource="#amis"/>
            <thes:hasScopeNote>
                  <rdf:Description>
                        <rdf:type
rdf:resource="http://www.itd.clrc.ac.uk/Projects/Limber/thesaurus#ScopeNote'
                        <thes:inLanguageOf
rdf:resource="http://www.itd.clrc.ac.uk/Projects/Limber/ISO639#fr"/>
                        <thes:hasTypeOf
rdf:resource="http://www.itd.clrc.ac.uk/Projects/Limber/thesaurus#Translati
                        <rdf:value>Ce mot en Anglais n'est que utilise pou
relations Platonique.</rdf:value>
                  </rdf:Description>
            </thes:hasScopeNote>
            <thes:BroaderConcept rdf:resource="#983"/>
            <thes:NarrowerConcept rdf:resource="#329"/>
            <thes:RelatedConcept rdf:resource="#243"/>
            <thes:TopConcept rdf:resource="#345"/>
      </rdf:Description>
      <rdf:Description rdf:ID="amis">
```

```
            <rdf:type
rdf:resource="http://www.itd.clrc.ac.uk/Projects/Limber/thesaurus#Term"/>
            <thes:inLanguageOf
rdf:resource="http://www.itd.clrc.ac.uk/Projects/Limber/ISO639#fr"/>
            <rdf:value>amis</rdf:value>
      </rdf:Description>
      <rdf:Description rdf:ID="983">
            <rdf:type
rdf:resource="http://www.itd.clrc.ac.uk/Projects/Limber/thesaurus#Concept"/>
            <thes:ClassificationCode>983</thes:ClassificationCode>
            <thes:NarrowerConcept rdf:resource="#620"/>
            <thes:TopConcept rdf:resource="#345"/>
      </rdf:Description>
      <rdf:Description rdf:ID="329">
            <rdf:type
rdf:resource="http://www.itd.clrc.ac.uk/Projects/Limber/thesaurus#Concept"/>
            <thes:ClassificationCode>329</thes:ClassificationCode>
            <thes:BroaderConcept rdf:resource="#620"/>
            <thes:TopConcept rdf:resource="#345"/>
      </rdf:Description>
      <rdf:Description rdf:ID="243">
            <rdf:type
rdf:resource="http://www.itd.clrc.ac.uk/Projects/Limber/thesaurus#Concept"/>
            <thes:ClassificationCode>243</thes:ClassificationCode>
            <thes:RelatedConcept rdf:resource="#620"/>
      </rdf:Description>
      <rdf:Description rdf:ID="345">
            <rdf:type
rdf:resource="http://www.itd.clrc.ac.uk/Projects/Limber/thesaurus#Concept"/>
            <thes:ClassificationCode>345</thes:ClassificationCode>
            <thes:hasScopeNote>
                  <rdf:Description>
                        <rdf:type
rdf:resource="http://www.itd.clrc.ac.uk/Projects/Limber/thesaurus#ScopeNote"
                        <thes:inLanguageOf
rdf:resource="http://www.itd.clrc.ac.uk/Projects/Limber/ISO639#en"/>
                        <thes:hasTypeOf
rdf:resource="http://www.itd.clrc.ac.uk/Projects/Limber/thesaurus#Hierarchy"
                        <rdf:value>This hierarchy is ...</rdf:value>
                  </rdf:Description>
            </thes:hasScopeNote>
      </rdf:Description>
</rdf:RDF>
```

9. STANDARD THESAURUS API ACCESS FUNCTIONS

The standard thesaurus API access functions would be expected to provide the follc functions following the 6 relations using 2 or three letter acronyms used in th standards for mono-lingual (ISO 2788) and multi-lingual (BS 6723:1985) thes construction. Other functions are added to provide single relations used in the UN thesaurus, which can be obtained by sequences of the simpler relations, but which v be more efficiently performed within the thesaurus management system:

1. Get all terms related to a term X (across trees) - RT
2. Get all terms broader than a term X (within a tree) - BT
3. Get all terms narrower than a term X (within a tree) - NT
4. Get all synonyms for a term - UF
5. Get the root term for a term - TT
6. Get the scope node - SN - extended for the 3 types of scope node used i thesaurus model.
7. Get the whole Hierarchy
8. Get all terms containing word X
9. Get classification code for term X
10. Get term X in language Y
11. Get preferred term for X - USE

Users could explicitly select the source/target languages and relation (as well as the of course); or the source language could come from the user's profile (in the host sy but in this case the user should always be reminded about the consequences of hi settings.

An extreme case would be where the target language is chosen from the metadata particular data repository (so that the thesaurus is used to translate queries for part data resources, possibly completely behind the scenes with respect to the user). would require quite tight integration with the host system, and may not be possible the architecture.

CABINET OFFICE
Office of the e-Envoy
Leading the drive to get the UK online
delivering UK online

Design/selection criteria for software used to handle controlled vocabularies

Contents

1. Introduction and background 2
2. Definitions and conventions 2
3. User support interface 4
4. Meta-tagging tools 6
5. Mapping tools 9
6. Vocabulary management software 10
7. References and further reading 13

1. Introduction and background

1.1 Metadata standard

This document has been prepared in the context of practical implementation of the Government Category List (GCL), a controlled vocabulary that forms part of the e-Government Metadata Standard (e-GMS). The e-GMS requires every electronic resource to have Subject metadata drawn from the GCL (assigned either individually or to a file or folder of resources of which this item forms a part). Specifically, at least one GCL category describing the main subject of the resource should be allocated to Subject.category.

1.2 Interlinked vocabularies

While the GCL is the principal controlled vocabulary under consideration, it is by no means the only one. The e-GMS recommends the use of additional controlled vocabularies for populating Subject.keyword. Furthermore, government departments and other public sector bodies are encouraged to develop their own vocabularies attuned to the needs of their own subject areas and anticipated audiences. Effective and efficient operation of the e-GMS will depend on careful integration of a set of vocabularies and the software tools for handling them, by every organisation committed to putting its electronic resources at the service of the public. The use of such tools is not compulsory, but reduces the level of operational errors and the costs of data handling.

1.3 The tools needed

The types of vocabulary for which software tools should provide support include thesauri conforming to ISO 2788, classification schemes of various types, subject heading lists, taxonomies (typically combining the hierarchical properties of classification schemes with the reciprocal relationships and other features of thesauri) and simple authority lists. The software elements needed to manipulate these include:

1. User support interface
2. Meta-tagging tools
3. Mapping tools
4. Vocabulary management tools.

With the exception of the fourth type of software, which may well stand alone, the others are typically components in a package that has many functions other than handling Subject vocabularies. The criteria described below should therefore be combined with a much longer list of criteria covering the full functionality of each type of tool required.

1.4 Structure of the present document

Since the scope of this document is limited to vocabulary functions, it has been constructed in a modular way. Sections 3, 4, 5 and 6 may be read independently of each other. Any of them could be lifted, adapted to the context of a particular application and complemented with other functions where necessary. Sections 1, 2 and 7, comprising introduction, definitions and references, are common to all the modules.

2. Definitions and conventions

The terminology associated with controlled vocabularies can be confusing, especially as customs vary depending on the type of vocabulary. For simplicity, in this document we shall apply standard thesaurus terminology throughout. For more background on the standard, see ISO 2788: *Guidelines for the establishment and development of monolingual thesauri.*

- **Controlled vocabulary.** A finite list of terms, each one having a prescribed meaning, designed for use in classifying or indexing documents, and for searching them.
- **Thesaurus.** A controlled vocabulary designed primarily to assist searching in postcoordinate information retrieval systems. The terms in a thesaurus are formally

organised to avoid ambiguities and to show hierarchical and associative relationships between them (see definitions of broader, narrower and related terms below). The purpose of a thesaurus is to guide both the indexer and the searcher to select the same term or combination of terms to represent a given concept.

- **Classification scheme**. A controlled vocabulary designed for classifying or categorising resources. The terms in a classification scheme are arranged in hierarchies, which are often well adapted to assist browsing.

- **Taxonomy**. A classification scheme designed for use with electronic media, and often incorporating thesaural features. The GCL is an example of a taxonomy.

- **Term**. A word or phrase used to designate a concept. Where more than one term in a controlled vocabulary can designate the same concept, one is chosen as a **preferred term** and the others are treated as **non-preferred terms**.

- **Preferred term.** A term used in a controlled vocabulary to represent a given concept. In the GCL the preferred terms are usually known as "categories".

- **Non-preferred term**. The synonym or quasi-synonym of a preferred term. Non-preferred terms are not themselves used for classifying or indexing resources, but are provided as entry points to help people find the most appropriate preferred terms. (For example, "Senior citizens *Use* Older people".) They are also known as lead-in terms.

- **Category**. A broad class or subject area that may be used for classifying information resources. The preferred terms in the GCL are usually known as categories.

- **Inter-term relationships**. These are links between terms, of the following three types:

    - **Equivalence** is the relationship between a preferred term and a corresponding non-preferred term. This relationship is reciprocal and directional.
    - A **Hierarchical** relationship is that between a Broader Term and a Narrower Term. It is reciprocal and directional.
    - An **Associative** relationship links two Related Terms. It is reciprocal but non-directional.

- **Broader term** (abbreviated **BT**). A preferred term with a meaning wider in scope than the term in question, and shown superordinate to the latter in a hierarchy.

- **Narrower term** (abbreviated **NT**). A preferred term with a meaning narrower in scope than the term in question, and shown subordinate to the latter in a hierarchy.

- **Top term** (abbreviated **TT**). A term which has no Broader Terms in the vocabulary.

- **Related term** (abbreviated **RT**). A preferred term which is associatively related to the term in question.

- **Siblings**. Two or more terms with the same immediate Broader Term.

- **Lead-in entry**. In a controlled vocabulary, an entry provided to guide a user from a non-preferred term to the corresponding preferred term.

3. User support interface

This is the interface mediating user access to information resources or services. (Although there may be common features, it is not the same as the interface for managing the vocabulary, described in Section 6. Even where functions are the same in principle, look and feel will probably be adapted to suit the different context.) We are assuming that it is designed to exploit a controlled vocabulary, such as the GCL. In most cases terms from that same controlled vocabulary will be present in the metadata of the resources presented. The criteria below will apply not just to the interface but also to the functionality lying behind it, assisting users to browse through the vocabulary and select headings for search and/or display of information resources.

3.1 Browse facilities

All of the vocabulary types enumerated above should be accessible via a browsing facility. This is particularly important for classification schemes or taxonomies such as the GCL, and applicable also to thesauri.

- The user must be able to see a list of the top terms in the vocabulary.

- Hierarchical browsing must be supported. That is to say, on selecting any preferred term, the user should be able to display any narrower terms under it (as well as cross-references and scope notes). Typically the display will also show the highest ranking of any resources retrieved by the term (See 3.3 below), or at least the total number of postings. Similarly on selecting one of the narrower terms, the narrower terms at the level below should be displayable, and so on to as many levels as are present in the vocabulary. As an optional alternative, all levels of hierarchy may be presented at one click on the top level term.

- During hierarchical browsing, the hierarchical route(s) connecting the term currently displayed to the top of the tree should be displayed. It should be possible for the user to navigate easily to any node of the hierarchical routes displayed, in order to view the corresponding resources and/or narrower terms etc. Similarly he should be able to return to the list of top terms.

- When a preferred term is selected, it should be possible to view the cross-references to any related terms in the vocabulary.

- When a cross-reference is selected, it should be possible to navigate directly to the display for the related term.

- When a term is selected, it should be possible to view any scope note associated with the term.

- A search capability should allow the user to search the whole of the vocabulary for any preferred term that contains his search word or phrase, or which has a lead-in entry containing the word (for example, a search for "jobs" should lead the user to the term "Unemployment and jobseeking" via its lead-in entry "Action teams for jobs *Use* Unemployment and jobseeking".) If the user's search term is not present in the vocabulary, the software should respond by showing him the nearest matches found.

- While a user is searching, careful design should ensure that he can easily tell whether he is searching the vocabulary itself, or the resources indexed with the vocabulary, or both.

3.2 Search facilities

While some classification schemes and taxonomies are designed to support searching, the GCL was designed primarily to support browsing. Therefore, the functions in this group may

be useful but are not vital for its operation. Search facilities are, however, essential for the operation of thesauri and subject heading lists. They are also vital for searching the full text of resources. System implementers should bear in mind that search terms give very different results depending on whether they are applied to controlled vocabulary metadata or to free text searching. Well designed systems allow advanced users to choose which of these modes they want to apply.

- Users should be able to enter search terms directly, or to identify them by browsing through the vocabulary using the functions described above. Novice users especially should not be required to have knowledge of search syntax, operators or query language.

- Optional system settings should be available for the default mode of searching, to determine whether selected terms should be applied only to metadata, or to the full text of resources (including their metadata).

- The interface for advanced users should support them in developing search statements that may combine more than one subject term and perhaps other elements such as date, creator, etc. Where Boolean logic is available, the interface should support selection of terms, combined with correct syntax, during the process of thesaurus browsing. The option of applying terms to free text or to metadata should be available.

- When a term is selected for searching, the user should be offered the option of searching on that term alone, or of extending the search to all of its narrower terms and indeed the complete hierarchy below the original term. When the latter option is chosen, the complete search statement comprises all of the terms, with OR logic between each one and the next.

3.3 Retrieval and display of resources

- The resources retrieved and displayed under a preferred term should be all those which have that term present in their subject metadata, identified as being drawn from the controlled vocabulary in question. (A particular term such as "Depression" or "Chips" may have different meanings in different vocabularies, but these may be distinguished if the syntax of the metadata clarifies which was the source vocabulary.) Optionally, resources may be retrieved which have the term present in their title or text but not their Subject metadata. It is usually best to present these resources separately, since they may be less relevant to the request.

- Different retrieval capabilities are needed to handle monohierarchical and polyhierarchical vocabularies, respectively. In the polyhierarchical case (the GCL is an example), each individual preferred term is unique, and has the same scope and meaning wherever it appears in the hierarchies. In monohierarchical vocabularies, a term may be used in two different places with two different meanings, dependent on the hierarchy of broader terms above it. It is important to select retrieval software that works appropriately for the type(s) of controlled vocabulary that will be used.

- For retrieval using a monohierarchical vocabulary, the resources selected may depend on the whole string associated with a preferred term. For example, the string "Meteorology/Depression" should not retrieve the same resources as the string "Health/Mental health/Depression". With this type of vocabulary, either the whole string must be present in the metadata, or a unique ID present in the metadata may be used to drive the selection.

- For retrieval using a polyhierarchical vocabulary such as the GCL, any string of broader terms associated with a preferred term should be ignored. For example, under the term "Common Agricultural Policy", all resources having this term present in their Subject metadata should be displayed, irrespective of whether the display is to be shown under a hierarchical route "Agriculture, environment and natural resources/Farming/Common Agricultural Policy" or under "International affairs and defence/European Union/Common Agricultural Policy".

3.4 Multilingual capabilities

The GCL is currently monolingual, but might eventually be enhanced by the addition of alternative languages. With a multilingual GCL, or for the handling of existing multilingual vocabularies, it should be possible for the user to switch to any of the languages supported. In this event, all the preferred terms, and any text describing relationships between the terms (such as "*See also*") should be presented in the desired language. The basic structure of the vocabulary, including all hierarchical and associative relationships, should correspond exactly across all the languages. Non-preferred terms and scope notes may or may not have corresponding entries in each of the languages.

3.5 Administrative capabilities

Before users can search or browse, the vocabulary has to be entered into the system; thereafter it will need maintenance. Entry by direct keying is possible, but this is recommended only for the simplest of systems, because of the potential for introducing errors. Here we shall assume that the vocabulary is managed in a separate system (see Section 6), from which periodic outputs are used to set up and maintain the vocabulary in the search/browse systems.

- The interface should be capable of importing a vocabulary comprising preferred terms, non-preferred terms, relationships between the terms (of the three types equivalence, hierarchical and associative), and scope notes. In the absence of a national or international standard for thesaurus record exchange formats, a format for exchange of GCL data will have to be agreed among stakeholders in the development of the e-GMS. (Alternatives based on ASCII and XML may also be necessary.)

- The interface should be capable of importing updates to the GCL and/or other controlled vocabulary. The updates should take the form of a set of complete new records exported from the vocabulary management system, plus a list of existing records to be deleted. The latter should be imported first. Then the new records are used to supplement or replace the records previously held.

- Following import or updating of the controlled vocabulary, the system should be capable of validating all the term relationships, using the validations set out in paragraph 6.2 (Inter-term relationships) below. Any errors should be reported.

- So that errors identified may be addressed, there must be a facility for hand correction of the terms, relationships and scope notes. (Note that this facility entails some risk and so must be restricted to a competent administrator.)

4. Meta-tagging tools

These are usually part of a content management system or web "publishing" software. An essential part of the publishing procedure is to ensure that all of the appropriate metadata elements are added. Here we shall deal only with Subject metadata.

Meta-tagging tools may optionally derive Subject metadata automatically from the full text of a resource and other metadata elements. More commonly, they provide an interface to support authors and administrators when browsing through the vocabulary in order to select and assign Subject metatags to a given resource or collection of resources, or to confirm those assigned automatically.

In the recommendations that follow, the person selecting the metatags will be referred to as the "indexer".

4.1 Administrative capabilities

Before indexers can select terms, the vocabulary has to be entered into the system; thereafter it will need maintenance. Entry by direct keying is possible for very simple systems, in which the indexer will be presented with a very small choice of Subject terms. But for more complex systems direct keying is a source of errors, and the tiniest discrepancy of spelling or spacing

can cause problems downstream. The recommendations below assume the general case, in which the vocabulary is maintained in a separate system. Periodic outputs from the vocabulary maintenance system are used to set up and maintain the meta-tagger files. In the event that the imported vocabulary is too complex for some indexers, administrators may wish to derive subsets from it, for selective presentation to particular indexers or groups of indexers.

- The meta-tagger should be capable of importing a vocabulary comprising preferred terms, non-preferred terms, relationships between the terms (of the three types equivalence, hierarchical and associative), and scope notes. In the absence of a national or international standard for thesaurus record exchange formats, a format for exchange of GCL data will have to be agreed among stakeholders in the development of the e-GMS.

- The meta-tagger should be capable of importing updates to the GCL and/or other controlled vocabulary. In the absence of a standard, the updates should take the form of a set of complete new records (which will supplement or replace the records previously held) plus a list of existing records to be deleted. The latter should be imported first.

- Following import or updating of the controlled vocabulary, the system should be capable of validating all the term relationships, using the validations set out in para 6 (Inter-term relationships) below. Any errors should be reported.

- So that errors identified may be addressed, there should be a facility for hand correction of the terms, relationships and scope notes. (Note that this facility entails some risk and so must be restricted to a competent administrator.)

- It should be possible to designate subsets of the total vocabulary for presentation to particular indexers. Subsets may comprise flat lists of terms, or structured lists that preserve the inter-term relationships of the original vocabulary. Either possibility should be supported.

4.2 Template properties and processing

- The metadata template for any resource should permit the entry of more than one Subject value.

- The metadata template for any resource originating from a public body in the UK should require the entry of at least one Subject value drawn from the GCL, either at the level of the resource itself or at the level of the file or folder of which it forms a part (as indicated in the Relation metadata of the same resource).

- When a GCL term is selected as a metadata value, it should be allocated to Subject.category. The complete hierarchical string associated with the term should NOT be added. The precise syntax will depend on whether HTML or XML is being used.

- When a term from a polyhierarchical thesaurus is selected, it should be allocated to Subject.keyword. Just as for the GCL, any hierarchical string associated with the term should be ignored for the purposes of meta-tagging. Appropriate syntax should be used to identify the source thesaurus.

- When the source vocabulary is a monohierarchical one allowing repetition of the same term with different meanings, then the metadata added must clarify the concept intended. One option is to include the whole hierarchical string that leads down to the required term. Alternatively, the term's unique identifier may be used, together with a means of locating the term in the source vocabulary (e.g. by using XML Namespace capabilities).

4.3 Automatic derivation of metadata

The detail of how the derivation is done is not described here, since typically it involves complex proprietary algorithms for the analysis of full text. Another factor in the derivation can be the context of who is entering the data, where, when, etc. For example, if all the

resources published by a particular creator are on a single subject, it becomes trivially simple to apply one Subject.category value by default to his/her entire output. But the general case is more complex and the following provisions apply.

- The system should allow default Subject values to be entered automatically in response to the presence or absence of a combination of other metadata values, such as Creator, Audience, Publisher, etc.

- If automatic methods are used to fill in Subject metadata, the results should be displayable for human review, correction if necessary, and confirmation. It should be possible to set the system up so that automatically supplied Subject values cannot be finally admitted without human approval.

- Automatic entry of Subject metadata is not recommended if monitoring shows that the results are unsatisfactory in more than 40% of cases, prior to human review.

4.4 Training an automatic meta-tagger

- For meta-taggers that rely on pre-training, the training interface should make it possible for a human operator to present to the meta-tagger a set of resources judged relevant to a particular preferred term in the vocabulary (the positive set). In the event that a set of non-relevant resources (the negative set) is also required, the interface should support easy transfer of resources from one set to the other, during the training process.

- When resources are switched from positive to negative set or vice versa, the interface should show what effect this has on any test resources entered into the system, and what characteristics (e.g. words or phrases present in the text) are causing the results noted. This will assist the operator in optimising the training set.

- If the meta-tagger uses a rule-base to derive entries, it should support editing of the rule-base alongside testing. After the rule-base for a given heading is entered, the interface should allow that rule-base to be run against the resource collection and show the results. Diagnostics for each resource matched to the heading should show what caused it to be selected.

4.5 Human selection/review support

- When adding metadata or reviewing automatically assigned metadata, the indexer should be enabled to see the full text of the resource in question.

- The indexer must be able to display a list of the top terms in the vocabulary.

- Hierarchical browsing must be possible. That is to say, on selecting any term, the indexer should be able to display any narrower terms under it (as well as related terms and scope note – see below). Similarly on selecting a narrower term, the narrower terms at the level below should be displayable, and so on to as many levels as are present in the vocabulary.

- During hierarchical browsing, the hierarchical route(s) connecting the term currently displayed to the top of the tree should be displayed. It should be possible for the indexer to navigate easily to any node of the hierarchical routes displayed, in order to view the corresponding narrower terms, related terms and scope note. Similarly he should be able to return to the list of top terms.

- When a term is selected for viewing, it should be possible to view any scope note associated with the term, and the cross-references to any related terms in the vocabulary.

- When the cross-reference to a related term is selected, it should be possible to navigate directly to the display for that term.

- A search capability should allow the indexer to search the whole of the vocabulary for any preferred term that contains his search word or phrase, or that has a lead-in entry containing the word (for example, a search for "jobs" should lead the indexer to the term "Unemployment and jobseeking" via its lead-in entry "Action teams for jobs *Use* Unemployment and jobseeking".)

- While browsing or searching, the indexer may at any time select one or more terms as Subject metadata values for the resource he is indexing. The software must permit easy transfer (for example, by double-clicking) of the selected terms to the metadata template of the resource. If he selects a non-preferred term, then the entry made in the template should be the corresponding preferred term.

- If an indexer enters a term directly to the Subject metadata, the software should check its validity against the controlled vocabulary applicable and complete any syntax requirements, such as identification of the source vocabulary. Any non-preferred term entered should be converted to the corresponding preferred term, and the indexer should be invited to confirm its appropriateness. If the term entered is not present in the source vocabulary, then it may be retained among the metadata, without any refinement, as an uncontrolled term.

- Online Help should be available to the indexer at any point in the selection and review process.

4.6 Bulk editing

- A bulk editing facility should be available in the event that (a) a change in the taxonomy requires retrospective correction of metadata already assigned to a collection of resources or (b) a batch of resources to be published all share the same Subject metadata values.

- The facility should be able to retrieve all the records to which an invalid term has been assigned to Subject metadata, and present them for human review and correction.

- The system should also allow automatic "Find and Replace" for all instances where one term in the metadata should be replaced by another (or by a combination of two or more terms).

5. Mapping tools

These are designed to infer suitable GCL categories and add them to Subject metadata, given the presence of terms from another vocabulary. Typically they form a component integrated into the meta-tagging tools described above, but are sufficiently complex to merit separate description.

The scenario we are assuming is that a given organisation uses its own specialist vocabulary for applying Subject metadata to the resources it publishes on its own network or website. In order to conform to the e-GMS, the same resources should also be tagged with at least one category (preferred term) from the GCL. Several ways of adding the appropriate GCL category(ies) are possible, for example an indexer may add them directly, or an automatic categoriser may derive them from the text of the resource. The option described here is that GCL categories are derived automatically by conversion from the specialist terms already assigned.

One of the simplest ways of deriving the GCL category(ies) is to employ a mapping table. The table shows, for each term present in the source vocabulary, what is the corresponding category or combination of categories in the GCL. During the meta-tagging process, as soon as Subject tags from the specialist vocabulary have been selected, a tool integrated with the meta-tagger looks up the mapping table and adds the corresponding GCL categories to the template. Other approaches are acceptable too, provided they achieve the result that (a) at least one valid GCL category is added to the subject metadata of each resource published, and (b) the category(ies) added do reflect the main subject of the resource.

Recommendations for the above scenario are as follows:

5.1 Mapping process and support

- As soon as one or more metadata terms from the specialist vocabulary have been confirmed, the mapping process should be initiated automatically. In some models it may be acceptable to perform this process in batch, say overnight. But in general real-time processing is highly desirable.

- From a mapping table, the GCL category(ies) corresponding to each specialist term should be added to Subject.category. Optionally, additional rules may be applied to supplement the GCL categories with others, or to reduce the number of GCL categories entered to a smaller number reflecting the most important subject aspects.

- Facilities should be available to let a human indexer review the GCL categories added, and if necessary correct them. GCL browsing and term selection should be available during the review process, just as during normal meta-tagging.

6. Vocabulary management software

Specialist software is needed to build and maintain the GCL and other vocabularies. Since specialist vocabularies can be remarkably varied, not all the possibilities can be covered here. We shall not discuss monohierarchical vocabularies, nor the management of widely used schemes such as the Dewey Decimal Classification (DDC), the Library of Congress Subject Headings (LCSH) or the Universal Decimal Classification (UDC), which have their own maintenance agencies. We shall, however, cover the requirements for handling the GCL itself, any vocabulary modelled on the GCL, and any thesaurus conforming to ISO 2788. The structure of these vocabularies is generally polyhierarchical, and typically they are managed in a database holding a record for every term (preferred or non-preferred). Each record shows relationships to other terms, as well as the term itself.

6.1 Size and character limitations

- There should be no limitations on the number of terms in the vocabulary which prevent it expanding to the size needed. (The GCL is unusually small, with currently about 400 preferred terms and about 1,300 non-preferred. Specialised thesauri may have 100,000 terms or more, and for multilingual thesauri this number must be multiplied by the number of languages present.)

- It should be possible to enter terms of 100 characters or even more, although terms will rarely exceed about 40 characters.

- There should be no limitation on the number of hierarchical levels admissible.

- The software should be able to handle the special characters sometimes found in terms, such as parentheses, apostrophe, ampersand etc. For scope notes, all the common punctuation marks should be admissible. For languages other than English, diacritical marks and non-ASCII characters may also be needed.

6.2 Inter-term relationships

- The software should not permit the coexistence of duplicate terms. Any duplicates entered should be submitted to the editor for correction, amalgamation, the addition of a qualifier, or other remedial action. Typographical differences such as italics or capitalisation should be ignored for the purposes of duplicate detection.

- The software must support the standard relationships BT/NT, RT, USE/UF

- The software must support the standard reciprocity requirements. That is to say, if Term A has a BT relationship with Term B, then Term B must have an NT relationship with Term A and vice versa; if Term C has an RT relationship with Term D, then Term D must have an RT relationship with Term C; if Term E has a USE relationship with Term F, then

Term F must have a UF relationship with Term E and vice versa. Preferably the software enters the reciprocal relationships automatically in response to the editor's insertion one way round; but as a minimum the software must issue a warning if any non-reciprocal relationship is present.

- When a term is amended or deleted, the change should be propagated automatically in all places where that term appears related to another (whether as BT, NT, RT, USE or UF). In the event of term deletion, all relationships to and from that term will thus be deleted.

- There should be no limitation on the number of relationships a given preferred term may have with others. Thus one term may have any number of BTs, NTs, RTs, UFs. (But note that certain combinations are inadmissible – see below.)

- It should be possible to set up user-defined reciprocal relationships, for example to distinguish between different types of BT/NT or different types of associative relationship.

- Validation checks must prevent the entry of inadmissible relationship combinations. If two terms already have one of the standard relationships, no other standard relationship between the same terms is admissible. If Term A has BT Term B, then none of the terms in the BT hierarchy above Term B should be admissible as BT, NT or RT of Term A. Non-preferred terms (that is to say, any term with a USE relationship to another term) may not have any BT, NT, RT or UF relationships.

- No relationship from a term to itself is admissible.

6.3 Multilingual relationships

- In the case of a multilingual thesaurus, each term must be linked to its corresponding term in each of the languages in the system. These linkages must be reciprocal.

- All of the BT/NT and RT/RT relationships present in any one of the languages should be replicated in each of the other languages.

- Any class codes or notation present should be available for application to all the languages.

- The Scope Notes, other types of note and USE/UF relationships in one language need not be matched in the other languages.

6.4 Term notes

- The software must support entry of a Scope Note (SN) associated with any term. Scope Notes are text notes of any length, typically clarifying the subject coverage of a given term.

- If a scope note makes reference to another term in the vocabulary, then the software should preferably support addition of a marker to the record for that term.

- The software should support the setting up of user-defined notes associated with any term, for example History Notes, Editorial Housekeeping Notes, etc.

6.5 Codes and notation

- It should be possible to associate at least one class code, number or other style of notation with any term. Preferably more than one type of notation should be supported.
- It should be possible to assign a unique identifier to each term.
- It should be possible to output the vocabulary using the sequence of each type of notation or identifier.

6.6 Node labels

- For the hierarchical or classified display of a thesaurus, it is sometimes useful to group terms under a "node label", especially when facet analysis is used to organise the terms. A node label indicates to which facet the terms under it belong, but is not a preferred term and has no place in the alphabetical display. Since node labels are an optional feature, the ability to enter and manipulate them is desirable but not essential.

6.7 Mapping management

- In the event that the vocabulary is a specialist one being used alongside another scheme such as the GCL, it may be convenient to maintain mapping correspondences within the database for the specialist vocabulary. For each term or group of terms, it should be possible to establish a mapping link to the corresponding GCL term. Such links will not be reciprocal, since the GCL master database is not managed within the same system.

- It must be possible to edit the mapping links each time the GCL is updated. A facility for bulk editing may prove very useful in the event that all occurrences of a given GCL term are to be changed in the same way.

6.8 Data import/export

- There should be a mechanism for bulk import of datasets from existing vocabularies, comprising terms, scope notes and standard relationships between the terms. All of the mentioned features should be retained after import.

- There must be a mechanism for producing reports or exporting the vocabulary, including all terms, scope notes, notation and standard relationships between the terms. Preferably it should also be possible to export subsets, such as non-preferred terms only, or preferred terms with their scope notes and NTs, etc.

- There must be a mechanism for exporting all records that have been amended after a given date. The option should be available of selecting only certain types of amendment, e.g. only new term records, or of including all records that have changed in any detail of the term itself or its relationships. It should also be possible to report all terms that have been deleted from a certain date.

- There must be a mechanism for outputting thesaurus displays, either in hard copy or on the screen. It should be possible to choose between a variety of sequences and layouts for the display. The main types are an alphabetical display, which lists each term in alphabetical order together with its notes and all relationships, and a systematic display. Systematic displays are generally hierarchical by top term or by class code. In systematic displays, it is often acceptable for all siblings below a parent term to be presented in alphabetical order, but the facility to force a different sequence is desirable.

6.9 Editorial navigation and support

- The editor must be able to go to the record for any term by direct entry of that term. Preferably he should only have to enter a portion of the term, and the software should assist his selection by identifying all terms that contain the portion.

- The editor should also be able to locate any term record by browsing via the term relationships. The browse facilities should preferably enable him to start off from a list of top terms, but in the absence of this he should at least be able to navigate from any term to the record for any of its NTs, BTs, RTs, USEs or UFs.
- The editorial interface should preferably allow viewing of a term's full hierarchical context, at the same time as the term record itself, with all its notes, codes and relationships presented for editing.

- It should be possible to move a term (together with all its NTs, NTs of those NTs, etc) easily from one hierarchical position to another, preferably using drag-and-drop. (NB This requires breaking one set of BT/NT relationships and making another. Any RT

relationships thus invalidated should be drawn to the editor's attention.)

- With a multilingual thesaurus, the editor should be able to switch his view easily from one language to any of the others.
- Editing capabilities should include standard word processing facilities, such as the ability to add, modify and delete characters without rekeying an entire field.
- When an editor takes steps to delete a term, the software should first seek confirmation that the deletion of that term is intended, before completing the action.
- When a relationship is to be established between two terms already present in the vocabulary, it should be possible to achieve this by navigation and selection rather than by having to key in the whole of a term already known to the system. (NB This is important to prevent errors, as well as to support efficiency.)

6.10 Editorial safeguards

- Editorial changes should be made in the first instance to a master database from which outputs are derived periodically for downstream processes such as meta-tagging or resource discovery applications.
- If more than one person is editing the master database simultaneously, an inbuilt mechanism must prevent simultaneous write access to the same records.
- Editorial changes from unauthorised persons must not be allowed.
- Preferably, the software should allow junior staff to make provisional changes, which are not finally admitted until approved by the principal Editor.
- It should be possible to vary the access privileges for a multilingual thesaurus, depending on which language version is being edited.
- For each editor, a roll-back (undo) function should allow progressive reversal of editorial changes he has entered.

6.11 General features

It would be too lengthy to list all the important criteria for software selection, such as robustness. But we shall list a few criteria that are significant for vocabulary-handling, even though not limited to this function.

- Good documentation of the system should be available
- Technical support should be available
- Training and online help screens should be available, and customisable to reflect the particular application.

7. References and further reading

7.1 Thesaurus construction

1. ISO 2788-1986 [=BS5723:1987] Documentation - Guidelines for the establishment and development of monolingual thesauri. Geneva: International Organization for Standardization; 1986.
2. Aitchison, Jean; Gilchrist, Alan, and Bawden, David. Thesaurus construction and use: a practical manual. 4th ed. London: Aslib; 2000. 218 pp.
3. Training courses on thesaurus and taxonomy development are regularly offered by TFPL (www.tfpl.com) and Aslib (www.aslib.com).

7.2 Documentation available on GovTalk

Name of document	Comment
GCL FAQs	The easiest place to start
Guide to Meta-tagging with the GCL	Advice for webmasters and authors of electronic resources when entering metadata
Specialised vocabularies and the GCL	Guidance for organisations using their own taxonomy or thesaurus as well as the GCL
Design/selection criteria for software used to handle controlled vocabularies	Helps with choosing software for any part of the implementation
e-Government Metadata Standard	Full details of all the metadata elements needed for interoperability in the public sector
GCL Maintenance Guide	Useful for the GCL editor and for developers of other category lists, thesauri, etc.
Implementing the GCL	Notes for system administrators, especially when using the plain text version to integrate GCL data into in-house systems
e-Government Metadata Standard (e-GMS)	Full details of all the metadata elements needed for interoperability in the public sector

Note: Comments on this document should be sent to govtalk@e-envoy.gsi.gov.uk

International Standard 5964

INTERNATIONAL ORGANIZATION FOR STANDARDIZATION•МЕЖДУНАРОДНАЯ ОРГАНИЗАЦИЯ ПО СТАНДАРТИЗАЦИИ•ORGANISATION INTERNATIONALE DE NORMALISATION

کتابخانه
انجمن اردو و تحقیقات صنعتی ایران

Documentation — Guidelines for the establishment and development of multilingual thesauri

Documentation — Principes directeurs pour l'établissement et le développement de thesaurus multilingues

First edition — 1985-02-15

Portal.acm.org

UDC 025.48

Ref. No. ISO 5964-1985 (E)

Descriptors : documentation, subject indexing, information retrieval, thesauri, multilingual thesauri, preparation, rules (instructions).

Price based on 61 pages

Contents

Page

0 Introduction 1

1 Scope and field of application 1

2 References 1

3 Definitions 1

4 General 2

5 Abbreviations and symbols 3

6 Vocabulary control 3

7 The establishment of a multilingual thesaurus: general problems 4

8 The establishment of a multilingual thesaurus: management decisions 5

9 The establishment of a multilingual thesaurus: language problems 6

10 Establishing equivalent terms in different languages 11

11 Other language problems 20

12 Relationships between terms in a multilingual thesaurus 21

13 Display of terms and relationships 24

14 Form and contents of a multilingual thesaurus 31

15 Organization of work 33

Annexes

A Symbolization of thesaural relationships 35

B Examples of displays 37

Documentation — Guidelines for the establishment and development of multilingual thesauri

0 Introduction

A trend towards the international exchange of information, fully supported by the UNISIST* programme of UNESCO, and exemplified by systems like the International information system for the agricultural sciences and technology (AGRIS) and the International Nuclear Information System (INIS), clearly calls for a higher commitment to multilingual cooperation. Information systems are expanding across language boundaries, leading to a notable increase in the provision of indexing and retrieval tools which are either language-independent (the Broad System of Ordering), or multilingual. Aids of this kind are essential if retrieval of documents indexed in more than one language is not to depend on the acquisition and use of a single, dominant language. Indexers or searchers should, where possible, be able to work in their mother tongues, or at least in a language with which they are already familiar. Within this context it is considered that multilingual thesauri have a significant part to play in improving the bibliographic control of literature on a global scale.

The standardization of procedures for the construction of a multilingual thesaurus is seen as a primary step in achieving compatibility between thesauri produced by indexing agencies using terms selected from different natural languages. The recording of these procedures will also enable indexers engaged in this task to benefit from the experience of others, and to work in a logical and consistent fashion, using recommended practices which have been established in the course of discussions at an international level.

1 Scope and field of application

1.1 The guidelines given in this International Standard should be used in conjunction with ISO 2788, and regarded as an extension of the scope of the monolingual guidelines. It is considered that the majority of procedures and recommendations contained in ISO 2788 are equally valid for a multilingual thesaurus. This applies particularly to general procedures concerning, for example, the forms of terms, the basic thesaural relationships, and management operations such as evaluation and maintenance. Except when it appears to be necessary, the procedures described in ISO 2788 are not repeated here, and it is therefore essential to refer to both of these International Standards when constructing a multilingual thesaurus.

1.2 These guidelines are restricted in scope to the problems of multilingualism which can arise during the construction of a "conventional" thesaurus, i.e. a thesaurus displaying terms selected from more than one natural language, these terms then constituting the vocabulary of a controlled indexing language. Throughout this International Standard, a distinction is made between *preferred terms* and *non-preferred terms* (see definitions in clause 3). These guidelines are not applicable to indexing languages in which concepts are expressed entirely as symbols (for example mathematical equations or chemical formulae), nor to systems which are based on the automatic analysis and searching of free text. It is considered, however, that a well-constructed multilingual thesaurus can play a significant part in improving retrieval from a free-text system which covers documents in more than one language.

1.3 Multilingual thesauri are relatively recent developments in the field of documentation, and it is inevitable, therefore, that the present guidelines should display certain limitations.

a) The examples used to illustrate problems encountered in the establishment of term equivalences have been drawn largely from the fields of science (including the social sciences) and technology. As far as possible, however, examples were chosen which illustrate general problems and procedures, i.e. those which should apply in any field of knowledge.

b) It is realized that the procedures described in these guidelines may not be entirely appropriate for all languages. The examples have been selected, for entirely pragmatic reasons, from three of the major languages, i.e. English, French and German, but this does not imply that these languages are regarded as dominant in the field of documentation. As far as possible the procedures considered here, together with their accompanying examples, relate to problems which may be encountered in any language.

* Intergovernmental programme for co-operation in the field of scientific and technological information.

Feedback would operate if, in response to these problems, the original German compound term is factored into its separate components, each expressed as a noun, i.e. "Bildung", "Gesetz" and "Lehrer", and if these are henceforth accepted as indexing terms in German and assigned to documents dealing with this subject. Translation into English and French could then be carried out on this new and simpler basis, i.e.

Gesetz	=	Law	=	Loi
Lehrer	=	Teachers	=	Enseignant
Bildung	=	Education	=	Formation

The German compound term "Lehrerbildungsgesetz" may still be retained in the German thesaurus if it is likely to be sought by users, but its status would be changed to that of a non-preferred term, and the user would be redirected to the combination of separate nouns which represents this complex concept, for example:

Lehrerbildungsgesetz BS LEHRER + BILDUNG + GESETZ

3.7 **indexing language**: A controlled set of terms selected from natural language and used to represent, in summary form, the subjects of documents.

NOTE — In a post-coordinate system these terms are used as "keywords" for retrieval purposes, usually without attempting to indicate their syntactical relationships. Syntactical relationships can be indicated in various ways in a pre-coordinated index, for example by printing terms in entries in an order which suggests their relative roles, and so allows the user to perceive the subject as a whole. Despite these differences, however, both kinds of system can be based on controlled vocabularies of terms displayed and organized in a thesaurus.

3.8 **indexing term**: The representation of a concept, preferably in the form of a noun or noun phrase.

NOTE — An indexing term can consist of more than one word, and is then known as a *compound term* (see 3.2). In a controlled indexing vocabulary, a term is designated either as a *preferred term* or as a *non-preferred term*.

3.9 **preferred term**: A term used consistently when indexing to represent a given concept; sometimes known as "descriptor".

3.10 **non-preferred term**: The synonym or quasi-synonym of a preferred term. A non-preferred term is not assigned to documents, but is provided as an entry point in a thesaurus or alphabetical index, the user being directed by an instruction (for example USE or SEE) to the appropriate preferred term; sometimes known as "non-descriptor".

3.11 **loan term**: A term in one language (the source language) which has been adopted without change in a second (or target) language.

Example:

German (source language)		**English** (target language)
KINDERGARTEN	=	KINDERGARTENS

3.12 **multilingual thesaurus**: A thesaurus (see 3.16) containing terms selected from more than one natural language. It displays not only the interrelationships between terms, but also equivalent terms in each of the languages covered.

3.13 **secondary language**: See definition of *dominant language* (3.3).

3.14 **source language (SL)**: That language which serves as a starting point when a preferred term is translated into its nearest equivalent term or terms in a second (or target) language.

3.15 **target language (TL)**: The language into which a term first encountered in another language (the source language, see 3.14) is translated.

3.16 **thesaurus**: The vocabulary of a controlled *indexing language* (see 3.7), formally organized so that the *a priori* relationships between concepts (for example as "broader" and "narrower") are made explicit.

4 General

No significance should be attached to the order in which terms in the various languages are set down in the examples, nor does the use of terms such as *source language* and *target language* imply that one or the other of these languages is dominant. The order of languages used in the examples, and the designation of one language as "source" and another as "target", varies from example to example, and depends upon the problem being illustrated. Where an appropriate order is not determined on technical grounds, alphabetical order (English, French, German) has been used. See definitions of source and target languages in 3.14 and 3.15.

6 Vocabulary control

6.1 Two principal means for achieving vocabulary control are employed in thesauri:

a) terms are deliberately restricted in scope to selected meanings. Unlike the terms in a dictionary, which may be accompanied by a number of different definitions reflecting common usage, each term in a thesaurus is generally restricted to whichever single meaning best serves the needs of an indexing system. The structure of a thesaurus, notably its display of hierarchical relationships, frequently indicates the intended meaning of a term. If this technique is not sufficiently explicit, a definition or scope note should be appended to the term. This should state the chosen meaning, and may also indicate other meanings which are recognized in natural language but which have been deliberately excluded for indexing purposes;

b) when the same concept can be expressed by two or more synonyms, one of these terms is usually selected as the preferred term (see 3.9) which is then used consistently in indexing. Reference to the preferred term should be made from any synonym which might also function as a user's access point. To assist the user of a printed thesaurus, it is recommended that preferred terms should be distinguished typographically from non-preferred terms.

6.2 Vocabulary control is also achieved through the application of rules or policies which determine, for example, the form of a term, for example its expression as a singular or plural, or the extent to which a compound term (see 3.2) should be either retained in its present form or factored into separate components, each of which is then expressed as a noun and used independently as an indexing term. These and other aspects of vocabulary control apply to all kinds of thesauri, whether monolingual or multilingual, and they are considered in greater detail in ISO 2788.

7 The establishment of a multilingual thesaurus: general problems

7.1 The establishment of any thesaurus, whether monolingual or multilingual, involves two different classes of problem which call for decisions:

a) management problems concerning, for example, the establishment of an updating policy, or the choice between alternative forms of display, etc.;

b) language problems which call for decisions on matters such as the forms of indexing terms (for example their representation as plurals or singulars), or the status of terms (i.e. preferred or non-preferred).

These general problems can usually be resolved by studying and choosing between the different recommended procedures described in a set of standard guidelines. Since these matters are covered already in ISO 2788, they will not be reconsidered here in detail.

7.2 The makers of a multilingual thesaurus face two additional kinds of problem which do not occur in monolingual work:

a) management problems, which call for decisions on, for example, the relative status of languages (i.e. the designation of a given language as the exchange language, the dominant language, or a secondary language), or the choice of a starting point for the work (for example translation of an existing thesaurus as opposed to *ab initio* construction);

b) language problems which call for the choice of an appropriate procedure when a term in one language expresses a concept which cannot be represented by an exactly equivalent term in one or more of the other languages.

7.3 These special problems associated with multilingual thesauri can be seen as extensions of the general problems considered in 7.1. The matter of status, for example, arises first as a general problem (i.e. the status of a term as "preferred" or "non-preferred"), and then re-occurs as a multilingual problem (i.e. the relative status of each language). The establishment of a multilingual thesaurus is, however, more complicated than this simple division of problems into two classes appears to indicate, due to the fact that some management decisions will directly affect the choice of procedures available to the indexer who is concerned with language problems. The possible extent of this interaction between two different classes of problem can be demonstrated by assuming, for example, that the editors of a new multilingual thesaurus have decided, as a matter of policy, to impose the following conditions:

a) the new multilingual thesaurus should be a translated version of an existing monolingual thesaurus;

b) the language of this source thesaurus should have the status of the exchange language;

c) feedback to this source thesaurus (see 3.6) is not allowed: that is to say, none of its terms or logical structure can be modified in response to linguistic or conceptual problems encountered in the other languages.

9 The establishment of a multilingual thesaurus: language problems

9.1 Introduction

9.1.1 The following assumptions are made in this and other clauses:

a) all languages in a multilingual thesaurus have equal status (see 3.4), whether or not one of these languages also functions as an exchange language;

b) either the *ab initio* method is adopted, or feedback is allowed if the work involves the translation and/or merging of one or more existing thesauri.

9.1.2 When languages have equal status, every preferred term in one of the languages should be matched by an equivalent preferred term in each of the other languages. It is not necessary to establish one-to-one equivalences between non-preferred terms, nor is this usually possible, since languages vary in the number of synonyms which express a given concept.

Examples:

a)

German		French		English
TELEFON BF Fernsprecher	=	TÉLÉPHONE	=	TELEPHONES
Fernsprecher BS TELEFON				

b)

German		French		English
FAHRSTUHL BF Aufzug	=	ASCENSEUR	=	LIFTS UF Elevators
Aufzug BS FAHRSTUHL				Elevators USE LIFTS

In case (a) a single term in French and in English is matched by two terms in German, i.e. a preferred term and a non-preferred term. In case (b) a single term in French is matched by two terms in German and two terms in English.

9.1.3 In the following clauses, the problems associated with the establishment of term equivalences are demonstrated, for the sake of simpler explanations, in only two languages. These examples have been chosen, however, to illustrate general procedures and practices which can be extended, by analogy, to situations involving more than two languages.

9.1.4 In the following examples the language possessing the term which gives rise to a particular translation problem is designated the source language (see 3.14). These problems usually occur when a term in a source language expresses a concept which has not been recognized by the users of the target language, so that the target language needs to be modified or extended to accommodate this "new" concept. The existence of such a problem could not have been recognized if translation had proceeded the other way round, i.e. if the "target language", as defined in 3.15, had served instead as the source of the translation, since this language lacks the term which introduces the problem. Consequently, the designation of a given language as "source" or "target" frequently varies, depending upon

a) the kind of problem being considered;

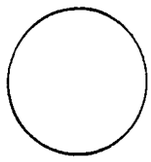
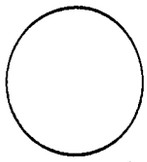
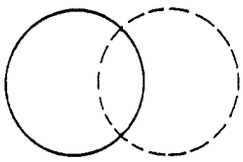
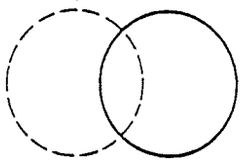
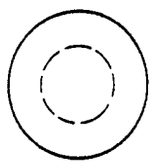
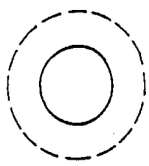
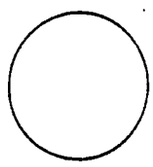
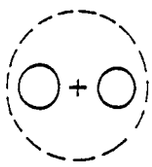
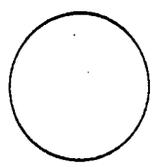
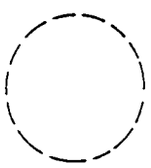
b) the stage reached in the construction of the thesaurus.

9.1.5 Experience gained in a number of international agencies indicates that acceptable equivalents between preferred terms can be established without difficulty in the majority of cases. This is sometimes as high as 90 %, although this figure may vary according to discipline, working procedures and language. Consequently, some of the procedures for dealing with non-equivalences described in the following clauses are likely to be applied in practice to only a small proportion of terms.

9.2 Degrees of equivalence and non-equivalence

9.2.1 Due to the nature of language itself, terms selected from more than one natural language vary in the extent to which they represent the same concepts. These variations can be regarded as forming a continuum, one end of which is represented by terms which can, for the practical purposes of indexing, be regarded as exact equivalents, further points being marked by various degrees of

Table 2 — Degrees of equivalence

Case	Source language	Target language
1 — Exact equivalence		
2 — Inexact equivalence		
3 — Partial equivalence		
4 — Single-to-multiple equivalence		
5 — Non-equivalence		

acceptable term exists

acceptable term does not exist

9.3.2 **Loan terms** (see 3.11)

9.3.2.1 The adoption of a loan term is usually necessary when it refers to a concept which is "native" to the users of the source language, and this concept is unlikely to arise independently within the community of target language users.

Examples:

a)	English DOLLARS	French DOLLAR	German DOLLAR
b)	French COGNAC	English COGNAC	German COGNAC

9.3.2.2 A loan term may also be adopted when its translation would call for a long definition or explanation which could not be used effectively as an indexing term in the target language.

Example:

English		German
TEENAGERS	=	TEENAGER D: Zwischen 13 und 19 Jahren

c) the invention of a neologism. This should be as concise as possible to encourage acceptance. These inventions may sometimes approximate to literal translations:

Examples:

English		French
STEAM CRACKING	=	VAPOCRAQUAGE NE: Craquage à la vapeur d'eau
TURBOFANS	=	TURBOSOUFFLANTE

— or they may, for cultural or linguistic reasons, express the concept from a different point of view:

Examples:

English		French
BULLDOZERS	=	BOUTEUR
SOFTWARE	=	LOGICIEL

10 Establishing equivalent terms in different languages

10.1 Exact equivalence (case 1, see 9.2.1)

Terms from different languages which refer to the same concept should be treated as exact equivalents. Exact equivalents can be morphologically related.

Example:

English		French		German
PHYSICS	=	PHYSIQUE	=	PHYSIK

— or they may be morphologically unrelated:

Example:

English		French		German
BLACKBIRDS	=	MERLE	=	AMSEL

— or they may appear to express the same concept from different viewpoints:

Example:

English		French
SOFT DRINKS	=	BOISSON NON ALCOOLISÉE

10.2 Inexact equivalence (case 2, see 9.2.1)

This situation covers terms which are generally regarded as denoting the same sets of objects or phenomena (for example they are frequently represented as equivalents in translation dictionaries), but the membership of these sets is slightly different.

Examples:

German		French
GEDECK	=	MENU

Solution: Terms which differ only in connotation should be treated, for indexing purposes, as exact equivalences.

10.3 Partial equivalence (case 3, see 9.3.1)

This situation covers terms which are generally regarded as referring to the same concept, but one of the terms strictly denotes a slightly broader or narrower concept.

Example 1:

English		German
SKIDDING	=	RUTSCHEN + SCHLEUDERN
		BF Rutschen
		BF Schleudern
		Rutschen
		BS RUTSCHEN + SCHLEUDERN
		Schleudern
		BS RUTSCHEN + SCHLEUDERN

Example 2:

English		French
FUELS	=	CARBURANT + COMBUSTIBLE
		EP Carburant
		EP Combustible
		Carburant
		EM CARBURANT
		+ COMBUSTIBLE
		Combustible
		EM CARBURANT
		+ COMBUSTIBLE

Solution B: Treat the source language term as a homograph, adding qualifying phrases or symbols to distinguish between the more specific meanings recognized in the target language.

Example 1:

English		German
SKIDDING (forwards)	=	RUTSCHEN
SKIDDING (sideways)	=	SCHLEUDERN

Example 2:

English		French
FUELS (motors)	=	CARBURANT
FUELS (heating)	=	COMBUSTIBLE

NOTE — The qualifying phrase attached to a term in this situation forms an integral part of the term, and it therefore differs from a scope note. If this solution is adopted, the general concept represented by the original term in the source language, for example SKIDDING, needs to be indexed by a combination of the qualified terms, for example SKIDDING (forwards) + SKIDDING (sideways).

Solution C: Adopt a partial combination of solution A and B.

Example 1:

English		German
SKIDDING	=	RUTSCHEN + SCHLEUDERN
		D Zu benutzen, wenn ein Dokument sowohl Rutschen als auch Schleudern behandelt
NT SKIDDING (forwards)	=	UB RUTSCHEN
NT SKIDDING (sideways)	=	UB SCHLEUDERN
SKIDDING (forwards)	=	RUTSCHEN
BT SKIDDING		OB RUTSCHEN + SCHLEUDERN
SKIDDING (sideways)	=	SCHLEUDERN
BT SKIDDING		OB RUTSCHEN + SCHLEUDERN

10.4.2 Single-to-multiple term equivalence: Situation 2

A compound term in the source language (see 3.2) represents a concept which is expressed by two or more separate terms in the target language, and the source language term can be factored[1] syntactically into components which, re-expressed as nouns if necessary, are then exact equivalents to the existing terms in the target language.

Example 1:

English		French
SOLAR HEATING	=	{ CHAUFFAGE { ÉNERGIE SOLAIRE

Example 2:

German		English
SCHÄFERHUNDERZIEHUNG	=	{ SHEEPDOGS { TRAINING

Solution: Three solutions to this problem are reviewed below. None of these represents a preferred solution, i.e. the choice will vary in practice according to circumstances.

Solution A: Except when the meaning of the compound term in the source language would be lost or distorted by factoring, the compound term should be factored into its semantic components, and these should be re-expressed as nouns (see note on "Feedback" in 3.6). These components should henceforth be assigned as separate terms to documents dealing with the compound concept. If the compound term is likely to be sought by users of the source language it should still be retained in the source language, but it should then be designated as a non-preferred term.

Example 1:

English		French
HEATING	=	CHAUFFAGE
SOLAR ENERGY	=	ÉNERGIE SOLAIRE

Solar heating
USE HEATING + SOLAR ENERGY

Example 2:

German		English
SCHÄFERHUND	=	SHEEPDOGS
ERZIEHUNG	=	TRAINING

Schäferhunderziehung
BS SCHÄFERHUND + ERZIEHUNG

Solution B: Where the meaning of the compound term in the source language would be lost or distorted by factoring, or if the term is in common use and its structure conforms to rules which govern the factoring or non-factoring of compound terms in the indexing language, retain the compound term and treat it as equivalent to a combination of terms in the target language. The terms which are used in combination in the target language should also be entered, with their equivalents, in the thesaurus.

Example:

English		French
SOLAR HEATING	=	CHAUFFAGE + ÉNERGIE SOLAIRE
HEATING	=	CHAUFFAGE
SOLAR ENERGY	=	ÉNERGIE SOLAIRE

1) See clause on factoring in ISO 2788.

Example 1:

French		English
GROS BÉTAIL	=	CATTLE + HORSES
		SN Use this combination as an equivalent to the French term GROS BÉTAIL
TS BOEUF		NT CATTLE
TS CHEVAL		NT HORSES

Example 2:

German		English		French
SCHNECKE	=	SLUGS + SNAILS	=	ESCARGOT + LIMACE
		SN Use this combination as an equivalent to the German term SCHNECKE		NE Combinaison équivalente à l'allemand SCHNECKE
UB GEHÄUSESCHNECKE		NT SNAILS		TS ESCARGOT
UB NACKTSCHNECKE		NT SLUGS		TS LIMACE

NOTE — This solution is feasible only when the combination consists of a limited number of terms (i.e. four or fewer). If several terms are involved, the combination becomes too unwieldy for the purposes of indexing and retrieval.

Solution B: Assign the status of a non-preferred term to the name of the extra category in the source language, and treat this term as an equivalent to whichever term in the target language is usually regarded as closest in meaning. This may be either a broader or a narrower term.

Example 1 — assuming that "GROS BÉTAIL" is regarded as closer in meaning to "LIVESTOCK" than to either "CATTLE" or "HORSES".

French		English
BÉTAIL	=	LIVESTOCK
EP Gros bétail		
TS BOEUF		NT CATTLE
TS CHEVAL		NT HORSES
Gros bétail		
EM BÉTAIL		

Example 2 — assuming that "SCHNECKE" is regarded as referring more often to "GEHÄUSESCHNECKE" than to "NACKTSCHNECKE".

German		English		French
GASTROPODE	=	GASTROPODA	=	GASTÉROPODES
UB GEHÄUSESCHNECKE		NT SNAILS		TS ESCARGOT
UB NACKTSCHNECKE		NT SLUGS		TS LIMACE
GEHÄUSESCHNECKE	=	SNAILS	=	ESCARGOT
BF Schnecke				
OB GASTROPODE		BT GASTROPODA		TG GASTÉROPODES
NACKTSCHNECKE	=	SLUGS	=	LIMACE
OB GASTROPODE		BT GASTROPODA		TG GASTÉROPODES
Schnecke				
BS GEHÄUSESCHNECKE				

NOTE — This solution allows the target language to retain all its own terms, but it also involves some loss of retrieval capability in the source language.

Solution C — Adopt the categorial name occurring in the source language as a loan term in the target language, adding a scope note to this term where its meaning is not self-evident.

Example:

German		English
BERUFSVERBOT	=	BERUFSVERBOT
		SN Alleged prohibition of certain classes of persons from official employment. Loan term adopted from German; used only in some political contexts.

Solution B: Coin a term in the target language which expresses the meaning of the source language term. This coined term may amount to a putative "translation" of the original term (see notes on coined terms in 9.3.3).

Example:

English		French
STEAM CRACKING	=	VAPOCRAQUAGE
		NE Craquage à la vapeur d'eau. Terme équivalent au terme anglais STEAM CRACKING

— or it may be an artificial invention:

Example:

English		French
BULLDOZERS	=	BOUTEUR
		NE Terme équivalent au terme anglais BULLDOZERS

10.6 Terms involving more than one kind of translation problem

10.6.1 In very rare circumstances it may be necessary to proceed in stages when translating a difficult term, and these stages might involve more than one of the categories of problems considered in 10.2 to 10.5 above, or the same kind of problem may be encountered more than once at different stages. It is impossible to illustrate all the various permutations of problems which can theoretically occur in these situations, nor can exact procedures be established which would be appropriate in every case. A general approach to these problems is therefore demonstrated below, using a single hypothetical example.

10.6.2 The problem chosen as a working example concerns the translation of the English term "teenagers" into French. It should be noted that the relative roles of the two languages, i.e. their functions as source or target language, occasionally change depending upon the stage reached in the translation.

Stage 1 — Establish the problem

English		French
TEENAGERS	=	?

Due to a difference between the English and French names for numbers, "teenagers" cannot be represented exactly by an existing term in French. At first sight, therefore, this appears to be a non-equivalence problem of the kind described in 10.5, and it could be handled by solution A, i.e. the adoption of "teenagers" as a loan term in French, or by solution B, i.e. the coining of a new term. For the sake of this demonstration, however, it will be assumed that neither of these solutions is acceptable, for example on the grounds that neither the loan term nor a coined term would be known to the users of the index. Study of a current bilingual dictionary suggests that the term "teenager" is generally translated as "adolescent". These terms can therefore be treated as examples of partial equivalents, and solution A in 10.3 could be applied.

Stage 2 — Treat as a partial equivalence problem (10.3); apply solution A

English		French
TEENAGERS	=	ADOLESCENT

It then becomes apparent that English already possesses a term "adolescents", which is an exact equivalent to the French term "adolescent", and the approach recommended in 10.1 would therefore give a more accurate rendering. Note that at this stage the roles of the languages have been reversed, i.e. French is now regarded temporarily as the source language.

11.1.3 Number

Thesauri in different languages frequently use different conventions when expressing nouns as singulars or plurals. In French and German, for example, singular forms, similar to those found in dictionaries, are generally preferred, whereas English language indexers usually select the plural or the singular according to rules associated with the kind of concept represented by the term. For example count nouns, frequently representing concrete entities, are expressed as plurals, while non-count nouns, which often represent abstract concepts, actions, etc., are expressed as singulars. Terms in a multilingual thesaurus should generally be expressed as singulars or plurals in accordance with the conventions recognized separately in each of the languages, especially when these are subject to national standards. There is no need to apply a single convention to all the languages. This means that non-count nouns would usually be expressed as singulars in all three languages in a thesaurus covering English, French and German.

Examples:

English		French		German
AIR	=	AIR	=	LUFT
FREEDOM	=	LIBERTÉ	=	FREIHEIT
DEAFNESS	=	SURDITÉ	=	TAUBHEIT

The treatment of count nouns, however, is likely to differ from one language to another.

English		French		German
ANIMALS	=	ANIMAL	=	TIER
CHILDREN	=	ENFANT	=	KIND
WATCHES	=	MONTRE	=	UHR

The conventions appropriate for each language should be explained in the introduction to a multilingual thesaurus, and should then be applied consistently. This recommendation has been applied to all the examples in this International Standard.

11.2 Homographs

11.2.1 These are distinguished as either:

a) **intra-language homographs**, i.e. a given form of term has more than one meaning within a single language;

b) **inter-language homographs**, i.e. a given form of term occurs in more than one language, but refers to a different concept in each language.

11.2.2 **Intra-language homographs:**

Example 1:

German		English
OTTER	=	ADDERS
OTTER	=	OTTERS

Example 2:

French		English
TOUR	=	TOWERS
TOUR	=	LATHES
TOUR	=	TOURS
TOUR	=	ROTATION

These homographs should be treated according to the recommendations in ISO 2788 or an equivalent national standard. Three solutions are reviewed below. None of these is regarded as the preferred solution, and the choice of solution will depend in practice on the special circumstances associated with each term.

11.2.3 Inter-language homographs:

Example:

English		French
BEAMS (radiation)	=	RAYON
RAYON	=	SOIE ARTIFICIELLE

Homographs of this kind are not likely to lead to confusion in a multilingual thesaurus except in the unlikely event that the editors decide to print terms from more than one language as a single alphabetical sequence, for example in the alphabetical index to parallel systematic listings. In that case, a sign to indicate language, for example an initial letter within parentheses, should be added to each of these terms.

Example:

RAYON (E)

RAYON (F)

11.3 Place names

11.3.1 Names of countries and geographical regions can vary from language to language.

Examples:

English		French		German
SPAIN	=	ESPAGNE	=	SPANIEN
THE HAGUE	=	LA HAYE	=	DEN HAAG

These terms are subject to the general recommendations concerning choice of term given in earlier clauses of this International Standard, i.e. that form of place name which is most familiar, or is officially recognized, within a given language community should be selected as the preferred term.

11.3.2 Variant forms of a place name sometimes coexist within a single language. These can arise for reasons such as the following:

— an "official" and a "popular" name are both in common use:

Example:

English		French		German
HOLLAND NETHERLANDS	=	HOLLANDE PAYS-BAS	=	HOLLAND NIEDERLANDE

— the original and a vernacular form coexist:

Example:

English
LEGHORN
LIVORNO

Solution: As a general rule the name which is most familiar to the users of the thesaurus should be designated as the preferred term. If all other things are equal, preference should be given to the official rather than the popular name.

Example 1:

English		French		German
NETHERLANDS UF Holland	=	PAYS-BAS EP Hollande	=	NIEDERLANDE BF Holland
Holland USE NETHERLANDS		Hollande EM PAYS-BAS		Holland BS NIEDERLANDE

11.4.3.3 Due to different transliteration and transcription systems, a name can sometimes appear with different spellings. The spelling which arises from the application of an International Standard should generally be selected as the preferred term. Other versions should then be treated as non-preferred terms, and references should be made from these to the designated preferred terms. When no suitable International Standard exists, or when an International Standard differs from a national standard, the source of the preferred spelling should be indicated.

Example:

English		French
CHEBYSHEV POLYNOMIALS UF Tchebyshev polynomials	=	POLYNÔME DE TCHEBYCHEV
Tchebyshev polynomials USE CHEBYSCHEV POLYNOMIALS		

11.5 Abbreviations and acronyms

11.5.1 As recommended in ISO 2788, abbreviations and acronyms should usually be designated as non-preferred terms, preference being given to the full form of the name.

Example:

English		French		German
DEOXYRIBONUCLEIC ACID UF DNA	=	ACIDE DÉSOXYRIBONUCLÉIQUE EP ADN	=	DESOXYRIBONUCLEINSÄURE BF DNS
DNA USE DEOXYRIBONUCLEIC ACID		ADN EM ACIDE DÉSOXYRIBONUCLÉIQUE		DNS BS DESOXYRIBONUCLEINSÄURE

11.5.2 Where an abbreviation is more widely known than the full form of the term, and provided that this is not likely to lead to confusion (for example as a homograph), the full form can be designated as the non-preferred term.

Example:

English		French		German
UNESCO UF United Nations Educational, Scientific and Cultural Organisation	=	UNESCO EP Organisation des Nations unies pour l'éducation, la science et la culture	=	UNESCO BF Organisation der Vereinten Nationen für Erziehung, Wissenschaft und Kultur
United Nations Educational, Scientific and Cultural Organisation USE UNESCO		Organisation des Nations unies pour l'éducation, la science et la culture EM UNESCO		Organisation der Vereinten Nationen für Erziehung, Wissenschaft und Kultur BS UNESCO

12 Relationships between terms in a multilingual thesaurus

12.1 The fundamental relationships described in ISO 2788 are regarded as language-independent and also culture-independent. They therefore apply equally to both monolingual and multilingual thesauri. Three kinds of basic relationship have been identified:

a) the equivalence relationship;

b) the hierarchical relationship;

c) the associative relationship.

Each of these is considered in turn below, with particular reference to the extent to which networks of terms linked by these relationships should correspond in the different language sections of a multilingual thesaurus.

vided that this term is regarded as useful by either of the language communities, then the "missing" term should be added to the hierarchy in the target language. If an acceptable equivalent does not exist, this should be treated in accordance with the recommendations in 10.3 to 10.5.

12.4 The associative relationship

12.4.1 As explained in ISO 2788, this basic relationship is the hardest to define. It refers to the link between two terms which do not form an equivalence set, nor can they be organized as a hierarchy, yet they are associated in common usage to such an extent that the user who refers to one of these terms in an index or thesaurus should be directed to the other term. It is necessary to exercise a measure of control over terms linked by this relationship. It can be stipulated, for example, that one of the terms should form a necessary component in any definition or explanation of the other.

Example 1:

English		French		German
BIRDS	=	OISEAU	=	VOGEL
RT ORNITHOLOGY		VA ORNITHOLOGIE		VB ORNITHOLOGIE
ORNITHOLOGY	=	ORNITHOLOGIE	=	ORNITHOLOGIE
RT BIRDS		VA OISEAU		VB VOGEL

Example 2:

English		French		German
INSECTS	=	INSECTE	=	INSEKT
RT INSECTICIDES		VA INSECTICIDE		VB INSEKTIZID
INSECTICIDES	=	INSECTICIDE	=	INSEKTIZID
RT INSECTS		VA INSECTE		VB INSEKT

In example 1, "Birds" is a necessary part of the definition of "Ornithology". In example 2, it is necessary to refer to "Insects" when explaining "Insecticides". Since this applies equally to all the languages, they should all display the same associative relationships.

12.4.2 Before an associative relationship which has been recognized in one language is transferred to another, it should be examined to determine the extent of its validity. If it appears to apply to only one group of language users, it should generally be excluded. In that case, the terms should also be re-considered to ensure that they do, in fact, refer to the same concept. Despite this injunction, a multilingual thesaurus should usually contain a richer variety of associative relationships than a monolingual thesaurus in the same field, since it will benefit from the viewpoints of different language users.

13 Display of terms and relationships

13.1 Introduction

13.1.1 Three forms of display are described in ISO 2788:

a) alphabetical arrangement, with scope notes, relationships, etc., indicated at each term;

b) systematic display, supported by an alphabetical index;

c) graphic display, supported by an alphabetical index.

Each of these basic kinds of display can also be used in a multilingual thesaurus, but the choice of method, and also the form of layout, is likely to be affected by the need, in a multilingual context, to indicate not only logical relationships between the terms in a given language, but also translations of the preferred terms in other languages.

13.1.2 This need to show translations of all preferred terms, and also to provide separate means of access to the thesaurus under each language, introduces two problems which do not occur in the construction of a monolingual thesaurus:

a) **Provision of alphabetical sequences in each language.** If two or more terms share a common relationship to a given preferred term (the set of all the narrower terms on the same level under a given higher term), these terms are usually arranged

13.2.3 An alphabetical multilingual thesaurus can be presented in two forms:

a) parallel columns showing not only translations, but also a full range of scope notes (if appropriate), related terms, etc., in each of the non-filing languages. This form is illustrated in figures 1A to 1C (annex B);

b) a single alphabetical sequence of terms in the filing language, with translations shown as part of the information attached to each term, but not in parallel columns. This form is illustrated in figures 2A to 2C (annex B).

13.2.4 From the reader's point of view, the presentation of equivalences in the form of parallel columns is probably easier to use, but it also demands more space. If the alternative form of display (see figures 2A to 2C) is adopted, further economies of space can be achieved by adopting techniques such as the following:

a) reducing the width of the column by presenting translations as a vertical list under each preferred term, rather than horizontally across the column;

b) showing translations of each of the filing terms, but without also showing translations of subordinate and other related terms.

Examples:

a) **English as filing language**

AEROPLANES
D: MOTORFLUGZEUG
F: AVION
SN fixed-wing, powered, heavier-than-air aircraft
BT FIXED-WING AIRCRAFT
NT FREIGHT AEROPLANES
JET AEROPLANES

b) **French as filing language**

AVION
D: MOTORFLUGZEUG
E: AEROPLANES
NE aéronef plus lourd que l'air, propulsé, à aile fixe
TG AÉRONEF À AILE FIXE
TS AVION CARGO
AVION À RÉACTION

c) **German as filing language**

MOTORFLUGZEUG
E: AEROPLANES
F: AVION
D nur für Luftfahrzeuge mit befestigten Flügeln, die schwerer als Luft sind
OB FLUGZEUG
UB FRACHTFLUGZEUG
DÜSENFLUGZEUG

13.3 Systematic display

13.3.1 As stated in ISO 2788, a systematic thesaurus should consist of two parts:

a) categories or hierarchies of terms arranged according to their meanings or logical relationships;

b) an alphabetical index which directs the user to the appropriate position in the systematic section.

The link between these two sections is provided by a system of address codes which are appended to each preferred term in the systematic section, and are then printed as references in the alphabetical index. These codes should have obvious filing values. They may consist of running numbers assigned to each preferred term (as in the examples of systematic display in annex B), or they may comprise a system of hierarchically expressive notation.

b) **the alphabetic section.** This contains a full range of scope notes and references between preferred and non-preferred terms. As a general rule the index gives references to broader, narrower and related terms, even when these are also shown in the graphic display.

13.4.3 Owing to the space constraints associated with tree structures and arrowgraphs, the alphabetical section necessarily contains a higher proportion of definitional and relational information, and it therefore functions as the main part in this kind of thesaurus. It may be as full as the alphabetical thesaurus described in 13.2, in which case the graphic display serves only to show, in diagrammatic form, hierarchical and other relationships which are not immediately obvious when using an alphabetical arrangement. In a thesaurus which covers a wide subject field, graphic displays may be limited to the preferred terms in selected core subjects. All terms appear, however, in the alphabetical section.

13.4.4 This is probably the hardest kind of thesaurus to adapt to multilingual use, owing to the space constraints mentioned above and the need to indicate inter-language equivalences. Various types of output are theoretically possible, depending upon management decisions concerning factors such as the following:

a) the choice of display method, for example tree structure or arrowgraph;

b) the position where equivalences are displayed, for example

1) only in the alphabetical sections, with separate monolingual graphic displays for each language;

2) in the graphic displays as well as in the alphabetical section.

The examples shown in annex B illustrate two relatively straightforward kinds of output :

Figures 5A to 5C: Tree structures in each language, with inter-language equivalences shown only in the alphabetical sections:

5A (i): English tree structure
5A (ii): Alphabetical section; English as filing language

5B (i): French tree structure
5B (ii): Alphabetical section; French as filing language

5C (i): German tree structure
5C (ii): Alphabetical section; German as filing language

Figures 6A to 6D: Arrowgraph containing preferred terms and equivalences in three languages, with separate alphabetical sections in each language:

6A: Arrowgraph

6B: English alphabetical section

6C: French alphabetical section

6D: German alphabetical section

14 Form and contents of a multilingual thesaurus

14.1 Owing to the range of displays available (see clause 13 and annex B), it is not possible to stipulate a standard form of layout in a printed thesaurus. The following parts should, however, be clearly distinguished:

a) title leaf (or leaves, if a separate leaf is produced in each language);

b) contents list(s);

c) introduction (see below);

d) systematic or graphic displays, where appropriate;

e) alphabetical section.

The title leaf or leaves should conform to the recommendations in ISO 1086.

15.2 Involvement of subject and language specialists

15.2.1 It is essential that editorial decisions concerning the terms of a given language should be fully endorsed by a native speaker of that language.

15.2.2 The need for help from language and/or subject specialists may arise at any time during the construction of a multilingual thesaurus. These specialists need not be engaged on a full-time basis, but the information scientists responsible for the work should be able to consult an appropriate specialist in circumstances such as the following:

a) defining terms and establishing their interrelationships;

b) selecting the preferred term when synonyms are encountered;

c) adopting loan terms, or coining new terms.

15.3 Initial stages of work, updating procedures, etc.

15.3.1 Avoidance of duplicate work

Before starting work on a multilingual thesaurus, every effort should be made to ascertain whether a thesaurus (monolingual or multilingual) already exists in the same subject field or in any cognate field. An enquiry should be addressed to one of the clearinghouses mentioned in 15.5.

15.3.2 Notification of intent

When an agency has decided to construct a multilingual thesaurus, notification of intent should be announced in an appropriate journal, for example *FID News Bulletin*.

15.3.3 Pilot run

It is recommended that the thesaurus should be tested in a pilot run prior to publication. A provisional version should be distributed to selected groups of users in each of the language communities. Any suggestions for changes to terms, their translations, and/or their interrelationships, should be considered by the editing committee before the text is approved for publication.

15.3.4 Updating procedures

These should be established, either on a continuous basis or at regular intervals, to deal with those changes which appear to be necessary in the light of experience. These changes may affect matters such as:

a) relationships between terms. Any change introduced into a hierarchy in one language must be equally valid when applied to the terms of another language;

b) the deletion of terms;

c) admission of terms which represent new concepts.

15.4 Other editorial matters

Reference should be made to ISO 2788 for full explanations of the following procedures which apply equally to monolingual and multilingual thesauri:

a) recording of terms;

b) term verification;

c) admission and deletion of terms;

d) the use of automatic data processing equipment.

Annex A

Symbolization of thesaural relationships

(This annex forms part of the standard.)

Abbreviations of the kind which have been used throughout this International Standard to represent basic thesaural relationships (for example BT, OB, TG and their equivalents in other languages) are frequently encountered in published thesauri, and they have acquired, through common usage, the status of "standard" conventions.

It is recognized, however, that these conventions represent abbreviations of terms selected from particular natural languages, and to that extent they are language-dependent. Some agencies might prefer more neutral and language-independent symbols which can be used in every section of a multilingual thesaurus. ISO has therefore prepared a special range of symbols to express the same relationships, and also to identify other components of an entry in a thesaurus, for example notes, language indicators, combined terms, etc.

Some of the symbols in table 3 do not appear on a standard typewriter keyboard, and only occasionally on computer print chains. They can be produced without difficulty, however, in any hand-set or photocomposed thesaurus.

It is stressed that these symbols, and the conventions they replace, are intended to express relationships only in printed thesauri. Relationships between terms in thesauri which are held in machine-readable form (for example on disk or tape) can be expressed by any suitable combination of characters, provided that standard symbols or conventions are generated in the printed output.

Table 3 — Symbols

Symbol	Significance
→	**Equivalence relationship**[1]
	Symbol which precedes:
	a) the preferred term
=	b) the non-preferred term
	Hierarchical relationship
	Symbol which precedes:
<	a) the broader or superordinate term
>	b) the narrower or subordinate term
⟝	c) the broader term (partitive)
⟞	d) the narrower term (partitive)
—	**Associative relationship**
	Symbol which precedes each related term
+	**Conjunction**
	Symbol printed between terms to indicate that they should be used in combination to represent a complex concept
X[2]	**Language indicator**

1) The distinction between preferred and non-preferred terms should also be indicated typographically, for example preferred terms in bold face or upper case, non-preferred terms in standard face.

2) "X" indicates a language code which is followed by a colon and a term in the indicated language, for example:

D: (term in German)
E: (term in English)
F: (term in French)

These codes should conform to ISO/R 639.

Aerodynes USE HEAVIER-THAN-AIR AIRCRAFT		
AEROPLANES SN fixed-wing, powered, heavier-than-air aircraft BT FIXED-WING AIRCRAFT NT FREIGHT AEROPLANES JET AEROPLANES	AVION NE aéronef plus lourd que l'air, propulsé à aile fixe TG AÉRONEF À AILE FIXE TS AVION CARGO AVION À RÉACTION	MOTORFLUGZEUG D nur für Luftfahrzeuge mit befestigten Flügeln, die schwerer als Luft sind OB FLUGZEUG UB FRACHTFLUGZEUG DÜSENFLUGZEUG
Aerostats USE LIGHTER-THAN-AIR AIRCRAFT		
AIRCRAFT NT HEAVIER-THAN-AIR AIRCRAFT LIGHTER-THAN-AIR AIRCRAFT MILITARY AIRCRAFT	AÉRONEF TS AÉRONEF PLUS LOURD QUE L'AIR AÉROSTAT AÉRONEF MILITAIRE	LUFTFAHRZEUG UB LUFTFAHRZEUG SCHWERER ALS LUFT AEROSTAT MILITÄRLUFTFAHRZEUG
AIRSHIPS BT LIGHTER-THAN-AIR AIRCRAFT	DIRIGEABLE TG AÉROSTAT	LUFTSCHIFF OB AEROSTAT
AVIATION NT GLIDING RT HEAVIER-THAN-AIR AIRCRAFT	AVIATION TS VOL À VOILE VA AÉRONEF PLUS LOURD QUE L'AIR	AVIATIK BF Flugwesen UB GLEITFLUG VB LUFTFAHRZEUG SCHWERER ALS LUFT
Cargo USE FREIGHT		
FIXED-WING AIRCRAFT BT HEAVIER-THAN-AIR AIRCRAFT NT AEROPLANES GLIDERS	AÉRONEF À AILE FIXE TG AÉRONEF PLUS LOURD QUE L'AIR TS AVION PLANEUR	FLUGZEUG OB LUFTFAHRZEUG SCHWERER ALS LUFT UB MOTORFLUGZEUG SEGELFLUGZEUG
FREIGHT UF Cargo RT FREIGHT AEROPLANES	FRET EP Cargaison VA AVION CARGO	FRACHT BF Kargo Ladung VB FRACHTFLUGZEUG
FREIGHT AEROPLANES BT AEROPLANES RT FREIGHT	AVION CARGO TG AVION VA FRET	FRACHTFLUGZEUG OB MOTORFLUGZEUG VB FRACHT
GLIDERS BT FIXED-WING AIRCRAFT RT GLIDING	PLANEURS TG AÉRONEF À AILE FIXE VA VOL À VOILE	SEGELFLUGZEUG BF Gleitflugzeug OB FLUGZEUG VB GLEITFLUG
GLIDING BT AVIATION RT GLIDERS	VOL À VOILE TG AVIATION VA PLANEUR	GLEITFLUG OB AVIATIK VB SEGELFLUGZEUG
HEAVIER-THAN-AIR AIRCRAFT UF Aerodynes BT AIRCRAFT NT FIXED-WING AIRCRAFT HELICOPTERS RT AVIATION	AÉRONEF PLUS LOURD QUE L'AIR TG AÉRONEF TS AÉRONEF À AILE FIXE HÉLICOPTÈRE VA AVIATION	LUFTFAHRZEUG SCHWERER ALS LUFT OB LUFTFAHRZEUG UB FLUGZEUG HUBSCHRAUBER VB AVIATIK
HELICOPTERS BT HEAVIER-THAN-AIR AIRCRAFT	HÉLICOPTÈRE TG AÉRONEF PLUS LOURD QUE L'AIR	HUBSCHRAUBER OB LUFTFAHRZEUG SCHWERER ALS LUFT
JET AEROPLANES BT AEROPLANES	AVION À RÉACTION TG AVION	DÜSENFLUGZEUG OB MOTORFLUGZEUG
LIGHTER-THAN-AIR AIRCRAFT UF Aerostats BT AIRCRAFT NT AIRSHIPS	AÉROSTAT TG AÉRONEF TS DIRIGEABLE	AEROSTAT OB LUFTFAHRZEUG UB LUFTSCHIFF
MILITARY AIRCRAFT BT AIRCRAFT	AÉRONEF MILITAIRE TG AÉRONEF	MILITÄRLUFTFAHRZEUG VB LUFTFAHRZEUG

Figure 1A — Alphabetical arrangement (English as filing language)

AEROSTAT	LIGHTER-THAN-AIR AIRCRAFT	AÉROSTAT
OB LUFTFAHRZEUG	UF Aerostats	TG AÉRONEF
UB LUFTSCHIFF	BT AIRCRAFT	TS DIRIGEABLE
	NT AIRSHIPS	
AVIATIK	AVIATION	AVIATION
BF Flugwesen		
UB GLEITFLUG	NT GLIDING	TS VOL À VOILE
VB LUFTFAHRZEUG SCHWERER ALS LUFT	RT HEAVIER-THAN-AIR AIRCRAFT	VA AÉRONEF PLUS LOURD QUE L'AIR
DÜSENFLUGZEUG	JET AEROPLANES	AVION À RÉACTION
OB MOTORFLUGZEUG	BT AEROPLANES	TG AVION
Flugwesen BS AVIATIK		
FLUGZEUG	FIXED-WING AIRCRAFT	AÉRONEF À AILE FIXE
OB LUFTFAHRZEUG SCHWERER ALS LUFT	BT HEAVIER-THAN-AIR AIRCRAFT	TG AÉRONEF PLUS LOURD QUE L'AIR
UB MOTORFLUGZEUG SEGELFLUGZEUG	NT AEROPLANES GLIDERS	TS AVION PLANEUR
FRACHT	FREIGHT	FRET
BF Kargo Ladung	UF Cargo	EP Cargaison
VB FRACHTFLUGZEUG	RT FREIGHT AEROPLANES	VA AVION CARGO
FRACHTFLUGZEUG	FREIGHT AEROPLANES	AVION CARGO
OB MOTORFLUGZEUG	BT AEROPLANES	TG AVION
VB FRACHT	RT FREIGHT	VA FRET
GLEITFLUG	GLIDING	VOL À VOILE
OB AVIATIK	BT AVIATION	TG AVIATION
VB SEGELFLUGZEUG	RT GLIDERS	VA PLANEUR
Gleitflugzeug BS SEGELFLUGZEUG		
HUBSCHRAUBER	HELICOPTERS	HÉLICOPTÈRE
OB LUFTFAHRZEUG SCHWERER ALS LUFT	BT HEAVIER-THAN-AIR AIRCRAFT	TG AÉRONEF PLUS LOURD QUE L'AIR
Kargo BS FRACHT		
Ladung BS FRACHT		
LUFTFAHRZEUG	AIRCRAFT	AÉRONEF
UB AEROSTAT	NT LIGHTER-THAN-AIR AIRCRAFT	TS AÉROSTAT
LUFTFAHRZEUG SCHWERER ALS LUFT	HEAVIER-THAN-AIR AIRCRAFT	AÉRONEF PLUS LOURD QUE L'AIR
MILITÄRLUFTFAHRZEUG	MILITARY AIRCRAFT	AÉRONEF MILITAIRE
LUFTFAHRZEUG SCHWERER ALS LUFT	HEAVIER-THAN-AIR AIRCRAFT	AÉRONEF PLUS LOURD QUE L'AIR
OB LUFTFAHRZEUG	UF Aerodynes	
UB FLUGZEUG HUBSCHRAUBER	BT AIRCRAFT	TG AÉRONEF
VB AVIATIK	NT FIXED-WING AIRCRAFT HELICOPTERS	TS AÉRONEF À AILE FIXE HÉLICOPTÈRE
	RT AVIATION	VA AVIATION
LUFTSCHIFF	AIRSHIPS	DIRIGEABLE
OB AEROSTAT	BT LIGHTER-THAN-AIR AIRCRAFT	TG AÉROSTAT
MILITÄRLUFTFAHRZEUG	MILITARY AIRCRAFT	AÉRONEF MILITAIRE
OB LUFTFAHRZEUG	BT AIRCRAFT	TG AÉRONEF
MOTORFLUGZEUG	AEROPLANES	AVION
D nur für Luftfahrzeuge mit befestigten Flügeln, die schwerer als Luft sind	SN fixed-wing, powered, heavier-than-air aircraft	NE aéronef plus lourd que l'air, propulsé, à aile fixe
OB FLUGZEUG	BT FIXED-WING AIRCRAFT	TG AÉRONEF À AILE FIXE
OB DÜSENFLUGZEUG FRACHTFLUGZEUG	NT JET AEROPLANES FREIGHT AEROPLANES	TS AVION À RÉACTION AVION CARGO
SEGELFLUGZEUG	GLIDERS	PLANEUR
BF Gleitflugzeug		
OB FLUGZEUG	BT FIXED-WING AIRCRAFT	TG AÉRONEF À AILE FIXE
VB GLEITFLUG	RT GLIDING	TS VOL À VOILE

Figure 1C — Alphabetical arrangement (German as filing language)

AÉRONEF
D: LUFTFAHRZEUG
E: AIRCRAFT
TG VÉHICULE D: FAHRZEUG E: VEHICLES
TG AÉRONEF MILITAIRE D: MILITÄRLUFTFAHRZEUG E: MILITARY AIRCRAFT
AÉRONEF PLUS LOURD QUE L'AIR
D: LUFTFAHRZEUG SCHWERER ALS LUFT
E: HEAVIER-THAN-AIR AIRCRAFT
AÉROSTAT D: AEROSTAT E: LIGHTER-THAN-AIR AIRCRAFT

AÉRONEF À AILE FIXE
D: FLUGZEUG
E: FIXED-WING AIRCRAFT
TG AÉRONEF PLUS LOURD QUE L'AIR
D: LUFTFAHRZEUG SCHWERER ALS LUFT
E: HEAVIER-THAN-AIR AIRCRAFT
TS AVION D: MOTORFLUGZEUG
E: AEROPLANES
PLANEUR D: SEGELFLUGZEUG E: GLIDERS

AÉRONEF MILITAIRE
D: MILITÄRLUFTFAHRZEUG
E: MILITARY AIRCRAFT
TG AÉRONEF D: LUFTFAHRZEUG E:AIRCRAFT

AÉRONEF PLUS LOURD QUE L'AIR
D: LUFTFAHRZEUG SCHWERER ALS LUFT
E: HEAVIER-THAN-AIR AIRCRAFT
TG AÉRONEF D: LUFTFAHRZEUG E: AIRCRAFT
TS AÉRONEF À AILE FIXE D: FLUGZEUG
E: FIXED-WING AIRCRAFT
HÉLICOPTÈRE D: HUBSCHRAUBER
E: HELICOPTERS
VA AVION D: AVIATIK E: AVIATION

AÉROSTAT
D: AEROSTAT
E: LIGHTER-THAN-AIR AIRCRAFT
TG AÉRONEF D: LUFTFAHRZEUG E: AIRCRAFT
TS DIRIGEABLE D: LUFTSCHIFF E: AIRSHIPS

AVIATION
D: AVIATIK
E: AVIATION
TS VOL À VOILE D: GLEITFLUG
E: GLIDING
VA AÉRONEF PLUS LOURD QUE L'AIR
D: LUFTFAHRZEUG SCHWERER ALS LUFT
E: HEAVIER-THAN-AIR AIRCRAFT

AVION
D: MOTORFLUGZEUG
E: AEROPLANES
NE aéronef plus lourd que l'air, propulsé, à aile fixe
TG AÉRONEF À AILE FIXE D: FLUGZEUG
E: FIXED-WING AIRCRAFT
TS AVION À RÉACTION D: DÜSENFLUGZEUG
E: JET AEROPLANES
AVION CARGO D: FRACHTFLUGZEUG
E: FREIGHT AEROPLANES

AVION À RÉACTION
D: DÜSENFLUGZEUG
E: JET AEROPLANES
TG AVION D: MOTORFLUGZEUG
E:AEROPLANES

AVION CARGO
D: FRACHTFLUGZEUG
E: FREIGHT AEROPLANES
TG AVION D: MOTORFLUGZEUG
E: AEROPLANES
VA FRET D: FRACHT E: FREIGHT

Cargaison EM FRET

DIRIGEABLE
D: LUFTSCHIFF
E: AIRSHIPS
TG AÉROSTAT D: AEROSTAT
E: LIGHTER-THAN-AIR AIRCRAFT

FRET
D: FRACHT
E: FREIGHT
EP Cargaison
VA AVION CARGO D: FRACHTFLUGZEUG
E: FREIGHT AEROPLANES

HÉLICOPTÈRE
D: HUBSCHRAUBER
E: HELICOPTERS
TG AÉRONEF PLUS LOURD QUE L'AIR
D: LUFTFAHRZEUG SCHWERER ALS LUFT
E: HEAVIER-THAN-AIR AIRCRAFT

PLANEUR
D: SEGELFLUGZEUG
E: GLIDERS
TG AÉRONEF À AILE FIXE D: FLUGZEUG
E: FIXED-WING AIRCRAFT
VA VOL À VOILE D: GLEITFLUG E: GLIDING

VÉHICULE
D: FAHRZEUG
E: VEHICLES
TS AÉRONEF D: LUFTFAHRZEUG E: AIRCRAFT

VOL À VOILE
D: GLEITFLUG
E: GLIDING
TG AVIATION D: AVIATIK E: AVIATION
VA PLANEUR D: SEGELFLUGZEUG E: GLIDERS

Figure 28 — Alphabetical arrangement (French as filing language)

201	VEHICLES	VÉHICULES	FAHRZEUG
202	AIRCRAFT	AÉRONEF	LUFTFAHRZEUG
203	HEAVIER-THAN-AIR AIRCRAFT	AÉRONEF PLUS LOURD QUE L'AIR	LUFTFAHRZEUG SCHWERER ALS LUFT
	UF Aerodynes		
	RT AVIATION 271	**VA** AVIATION	**VB** AVIATIK
204	FIXED-WING AIRCRAFT	AÉRONEF À AILE FIXE	FLUGZEUG
205	AEROPLANES	AVION	MOTORFLUGZEUG
	SN fixed-wing, powered, heavier-than-air aircraft	**NE** aéronef plus lourd que l'air, propulsé, à aile fixe	**D** nur für Luftfahrzeuge mit befestigten Flügeln, die schwerer als Luft sind
206	FREIGHT AEROPLANES	AVION CARGO	FRACHTFLUGZEUG
	RT FREIGHT 292	**VA** FRET	**VB** FRACHT
207	JET AEROPLANES	AVION À RÉACTION	DÜSENFLUGZEUG
208	GLIDERS	PLANEUR	SEGELFLUGZEUG
	RT GLIDING 275	**VA** VOL À VOILE	**BF** Gleitflugzeug **VB** GLEITFLUG
209	HELICOPTERS	HÉLICOPTÈRE	HUBSCHRAUBER
210	LIGHTER-THAN-AIR AIRCRAFT	AÉROSTAT	AEROSTAT
	UF Aerostats		
211	AIRSHIPS	DIRIGEABLE	LUFTSCHIFF
212	MILITARY AIRCRAFT	AÉRONEF MILITAIRE	MILITÄRLUFTFAHRZEUG

Figure 3A(i) — Systematic section (English as filing language)

401	VÉHICULE	VEHICLES	FAHRZEUG
402	AÉRONEF	AIRCRAFT	LUFTFAHRZEUG
403	AÉRONEF MILITAIRE	MILITARY AIRCRAFT	MILITÄRLUFTFAHRZEUG
404	AÉRONEF PLUS LOURD QUE L'AIR	HEAVIER-THAN-AIR AIRCRAFT	LUFTFAHRZEUG SCHWERER ALS LUFT
		UF Aerodynes	
	VA AVIATION 471	**RT** AVIATION	**VB** AVIATIK
405	AÉRONEF À AILE FIXE	FIXED-WING AIRCRAFT	FLUGZEUG
406	AVION	AEROPLANES	MOTORFLUGZEUG
	NE aéronef plus lourd que l'air, propulsé, à aile fixe	**SN** fixed-wing, powered, heavier-than-air aircraft	**D** nur für Luftfahrzeuge mit befestigten Flügeln, die schwerer als Luft sind
407	AVION À RÉACTION	JET AEROPLANES	DÜSENFLUGZEUG
408	AVION CARGO	FREIGHT AEROPLANES	FRACHTFLUGZEUG
	VA FRET 492	**RT** FREIGHT	**VB** FRACHT
409	PLANEUR	GLIDERS	SEGELFLUGZEUG
			BF Gleitflugzeug
	VA VOL À VOILE 475	**RT** GLIDING	**VB** GLEITFLUG
410	HÉLICOPTÈRE	HELICOPTERS	HUBSCHRAUBER
411	AÉROSTAT	LIGHTER-THAN-AIR AIRCRAFT	AEROSTAT
		UF Aerostats	
412	DIRIGEABLE	AIRSHIPS	LUFTSCHIFF

Figure 3B(i) — Systematic section (French as filing language)

601	FAHRZEUG	VEHICLES	VÉHICULES
602	LUFTFAHRZEUG	AIRCRAFT	AÉRONEF
603	AEROSTAT	LIGHTER-THAN-AIR AIRCRAFT	AÉROSTAT
		UF Aerostats	
604	LUFTSCHIFF	AIRSHIPS	DIRIGEABLE
605	LUFTFAHRZEUG SCHWERER ALS LUFT	HEAVIER-THAN-AIR AIRCRAFT	AÉRONEF PLUS LOURD QUE L'AIR
		UF Aerodynes	
	VB AVIATIK 671	**RT** AVIATION	VA AVIATION
	FLUGZEUG	FIXED-WING AIRCRAFT	AÉRONEF À AILE FIXE
	MOTORFLUGZEUG	AEROPLANES	AVION
	D nur für Luftfahrzeuge mit befestigten Flügeln, die schwerer als Luft sind	**SN** fixed-wing, powered, heavier-than-air aircraft	**NE** aéronef plus lourd que l'air, propulsé, à aile fixe
608	DÜSENFLUGZEUG	JET AEROPLANES	AVION À RÉACTION
609	FRACHTFLUGZEUG	FREIGHT AEROPLANES	AVION CARGO
	VB FRACHT 692	RT FREIGHT	VA FRET
610	SEGELFLUGZEUG	GLIDERS	PLANEUR
	BF Gleitflugzeug		
	VB GLEITFLUG 675	RT GLIDING	VA VOL À VOILE
611	HUBSCHRAUBER	HELICOPTERS	HÉLICOPTÈRE
612	MILITÄRLUFTFAHRZEUG	MILITARY AIRCRAFT	AÉRONEF MILITAIRE

Figure 3C(i) — Systematic section (German as filing language)

201 VEHICLES
202 AIRCRAFT
203 HEAVIER-THAN-AIR AIRCRAFT
UF Aerodynes
RT AVIATION 271
204 FIXED-WING AIRCRAFT
205 AEROPLANES
SN fixed-wing, powered, heavier-than-air aircraft
206 FREIGHT AEROPLANES
RT FREIGHT 292
207 JET AEROPLANES
208 GLIDERS
RT GLIDING 275
209 HELICOPTERS
210 LIGHTER-THAN-AIR AIRCRAFT
UF Aerostats
211 AIRSHIPS
212 MILITARY AIRCRAFT

Figure 4A(i) — Systematic section
(English as filing language)

Aerodynes **USE** HEAVIER-THAN-AIR AIRCRAFT
AEROPLANES 205
D: MOTORFLUGZEUG
F: AVION
SN fixed-wing, powered, heavier-than-air aircraft
Aerostats **USE** LIGHTER-THAN-AIR AIRCRAFT
AIRCRAFT 202
D: LUFTFAHRZEUG
F: AÉRONEF
AIRSHIPS 211
D: LUFTSCHIFF
F: DIRIGEABLE
AVIATION 271
D: AVIATIK
F: AVIATION
RT HEAVIER-THAN-AIR AIRCRAFT
Cargo **USE** FREIGHT
FIXED-WING AIRCRAFT 204
D: FLUGZEUG
F: AÉRONEF À AILE FIXE
FREIGHT 292
D: FRACHT
F: FRET
UF Cargo
RT FREIGHT AEROPLANES
FREIGHT AEROPLANES 206
D: FRACHTFLUGZEUG
F: AVION CARGO
RT FREIGHT
GLIDERS 208
D: SEGELFLUGZEUG
F: PLANEUR
RT GLIDING
GLIDING 275
D: GLEITFLUG
F: VOL À VOILE
RT GLIDERS
HEAVIER-THAN-AIR AIRCRAFT 203
D: LUFTFAHRZEUG SCHWERER ALS LUFT
F: AÉRONEF PLUS LOURD QUE L'AIR
UF Aerodynes
RT AVIATION
HELICOPTERS 209
D: HUBSCHRAUBER
F: HÉLICOPTÈRE
JET AEROPLANES 207
D: DÜSENFLUGZEUG
F: AVION À RÉACTION
LIGHTER-THAN-AIR AIRCRAFT
D: AEROSTAT
F: AÉROSTAT
UF Aerostats
MILITARY AIRCRAFT 212
D: MILITÄRLUFTFAHRZEUG
F: AÉRONEF MILITAIRE
VEHICLES 201
D: FAHRZEUG
F: VÉHICULE

Figure 4A(ii) — English alphabetical index

601 FAHRZEUG
602 LUFTFAHRZEUG
603 AEROSTAT
604 LUFTSCHIFF
605 LUFTFAHRZEUG SCHWERER ALS LUFT
VB AVIATIK 671
606 FLUGZEUG
607 MOTORFLUGZEUG
D nur für Luftfahrzeuge mit befestigten Flügeln, die schwerer als Luft sind
608 DÜSENFLUGZEUG
609 FRACHTFLUGZEUG
VB FRACHT 692
610 SEGELFLUGZEUG
VB GLEITFLUG 675
611 HUBSCHRAUBER
612 MILITÄRLUFTFAHRZEUG

Figure 4C(i) — Systematic section (German as filing language)

AEROSTAT 603
E: LIGHTER-THAN-AIR AIRCRAFT
F: AÉROSTAT
AVIATIK 671
E: AVIATION
F: AVIATION
BF Flugwesen
VB LUFTFAHRZEUG SCHWERER ALS LUFT
DÜSENFLUGZEUG 608
E: JET AEROPLANES
F: AVION À RÉACTION
FAHRZEUG 601
E: VEHICLES
F: VÉHICULES
Flugwesen BS AVIATIK
FLUGZEUG 606
E: FIXED-WING AIRCRAFT
F: AÉRONEF À AILE FIXE
FRACHT 692
E: FREIGHT
F: FRET
BF Kargo
Ladung
VB FRACHTFLUGZEUG
FRACHTFLUGZEUG 609
E: FREIGHT AEROPLANES
F: AVION CARGO
VB FRACHT
GLEITFLUG 675
E: GLIDING
F: VOL À VOILE
VB SEGELFLUGZEUG
Gleitflugzeug BS SEGELFLUGZEUG
HUBSCHRAUBER 611
E: HELICOPTERS
F: HÉLICOPTÈRE
Kargo BS FRACHT
Ladung BS FRACHT
LUFTFAHRZEUG 602
E: AIRCRAFT
F: AÉRONEF
LUFTFAHRZEUG SCHWERER ALS LUFT 605
E: HEAVIER-THAN-AIR AIRCRAFT
F: AÉRONEF PLUS LOURD QUE L'AIR
VB AVIATIK
LUFTSCHIFF 604
E: AIRSHIPS
F: DIRIGEABLE
MILITÄRLUFTFAHRZEUG 612
E: MILITARY AIRCRAFT
F: AÉRONEF MILITAIRE
MOTORFLUGZEUG 607
E: AEROPLANES
F: AVION
D nur für Luftfahrzeuge mit befestigten Flügeln, die schwerer als Luft sind
SEGELFLUGZEUG 610
E: GLIDERS
F: PLANEUR
BF Gleitflugzeug
VB GLEITFLUG

Figure 4C(ii) — German alphabetical index

Aerodynes USE HEAVIER-THAN-AIR AIRCRAFT

AEROPLANES T200
D: MOTORFLUGZEUG
F: AVION
SN fixed-wing, powered, heavier-than-air aircraft
BT FIXED-WING AIRCRAFT D: FLUGZEUG
F: AÉRONEF À AILE FIXE
NT FREIGHT AEROPLANES D: FRACHT-
FLUGZEUG F: AVION CARGO
JET AEROPLANES D: DÜSENFLUGZEUG
F: AVION À RÉACTION

Aerostats USE LIGHTER-THAN-AIR AIRCRAFT

AIRCRAFT T200
D: LUFTFAHRZEUG
F: AÉRONEF
BT VEHICLES D: FAHRZEUG F: VÉHICULE
NT HEAVIER-THAN-AIR AIRCRAFT
D: LUFTFAHRZEUG SCHWERER ALS LUFT
F: AÉRONEF PLUS LOURD QUE L'AIR
LIGHTER-THAN-AIR AIRCRAFT
D: AEROSTAT F: AÉROSTAT
MILITARY AIRCRAFT D: MILITÄRLUFT-
FAHRZEUG F: AÉRONEF MILITAIRE

AIRSHIPS T200
D: LUFTSCHIFF
F: DIRIGEABLE
BT LIGHTER-THAN-AIR AIRCRAFT
D: AEROSTAT F: AÉROSTAT

AVIATION R220
D: AVIATIK
F: AVIATION
NT GLIDING D: GLEITFLUG F: VOL À VOILE
RT HEAVIER-THAN-AIR AIRCRAFT
D: LUFTFAHRZEUG SCHWERER ALS LUFT
F: AÉRONEF PLUS LOURD QUE L'AIR

Cargo USE FREIGHT

FIXED-WING AIRCRAFT T200
D: FLUGZEUG
F: AÉRONEF À AILE FIXE
BT HEAVIER-THAN-AIR AIRCRAFT
D: LUFTFAHRZEUG SCHWERER ALS LUFT
F: AÉRONEF PLUS LOURD QUE L'AIR
NT AEROPLANES D: MOTORFLUGZEUG
F: AVION
GLIDERS D: SEGELFLUGZEUG F: PLANEUR

FREIGHT E240
D: FRACHT
F: FRET
UF Cargo
RT FREIGHT AEROPLANES D: FRACHT-
FLUGZEUG F: AVION CARGO

GLIDERS T200
D: SEGELFLUGZEUG
F: PLANEUR
BT FIXED-WING AIRCRAFT D: FLUGZEUG
F: AÉRONEF À AILE FIXE
RT GLIDING D: GLEITFLUG F: VOL À VOILE

GLIDING R220
D: GLEITFLUG
F: VOL À VOILE
BT AVIATION D: AVIATIK F: AVIATION
RT GLIDERS D: SEGELFLUGZEUG F: PLANEUR

HEAVIER-THAN-AIR AIRCRAFT T200
D: LUFTFAHRZEUG SCHWERER ALS LUFT
F: AÉRONEF PLUS LOURD QUE L'AIR
UF Aerodynes
BT AIRCRAFT D: LUFTFAHRZEUG F: AÉRONEF
NT FIXED-WING AIRCRAFT D: FLUGZEUG
F: AÉRONEF À AILE FIXE
HELICOPTERS D: HUBSCHRAUBER
F: HÉLICOPTÈRE
RT AVIATION D: AVIATIK F: AVIATION

HELICOPTERS T200
D: HUBSCHRAUBER
F: HÉLICOPTÈRE
BT HEAVIER-THAN-AIR AIRCRAFT
D: LUFTFAHRZEUG SCHWERER ALS LUFT
F: AÉRONEF PLUS LOURD QUE L'AIR

JET AEROPLANES T200
D: DÜSENFLUGZEUG
F: AVION À RÉACTION
BT AEROPLANES D: MOTORFLUGZEUG
F: AVION

LIGHTER-THAN-AIRCRAFT T200
D: AEROSTAT
F: AÉROSTAT
UF Aerostats
BT AIRCRAFT D: LUFTFAHRZEUG F: AÉRONEF
NT AIRSHIPS D: LUFTSCHIFF F: DIRIGEABLE

MILITARY AIRCRAFT T200
D: MILITÄRLUFTFAHRZEUG
F: AÉRONEF MILITAIRE
BT AIRCRAFT D: LUFTFAHRZEUG F: AÉRONEF

VEHICLES T200
D: FAHRZEUG
F: VÉHICULE
NT AIRCRAFT D: LUFTFAHRZEUG F: AÉRONEF

Figure 5A(ii) — Alphabetical section (English as filing language)

AÉRONEF T400
D: LUFTFAHRZEUG
E: AIRCRAFT
TG VÉHICULE D: FAHRZEUG E: VEHICLES
TS AÉRONEF MILITAIRE D: MILITÄRLUFTFAHRZEUG E: MILITARY AIRCRAFT
AÉRONEF PLUS LOURD QUE L'AIR
D: LUFTFAHRZEUG SCHWERER ALS LUFT
E: HEAVIER-THAN-AIR AIRCRAFT
AÉROSTAT D: AEROSTAT
E: HEAVIER-THAN-AIR AIRCRAFT

AÉRONEF À AILE FIXE T400
D: FLUGZEUG
E: FIXED-WING AIRCRAFT
TG AÉRONEF PLUS LOURD QUE L'AIR
D: LUFTFAHRZEUG SCHWERER ALS LUFT
E: HEAVIER-THAN-AIR AIRCRAFT
TS AVION D: MOTORFLUGZEUG
E: AEROPLANES
PLANEUR D: SEGELFLUGZEUG E: GLIDERS

AÉRONEF MILITAIRE T400
D: MILITÄRLUFTFAHRZEUG
E: MILITARY AIRCRAFT
TG AÉRONEF D: LUFTFAHRZEUG E: AIRCRAFT

AÉRONEF PLUS LOURD QUE L'AIR T400
D: LUFTFAHRZEUG SCHWERER ALS LUFT
E: HEAVIER-THAN-AIR AIRCRAFT
TG AÉRONEF D: LUFTFAHRZEUG E: AIRCRAFT
TS AÉRONEF À AILE FIXE D: FLUGZEUG
E: FIXED-WING AIRCRAFT
HÉLICOPTÈRE D: HUBSCHRAUBER
E: HELICOPTERS
VA AVIATION D: AVIATIK E: AVIATION

AÉROSTAT T400
D: AEROSTAT
E: LIGHTER-THAN-AIR AIRCRAFT
TG AÉRONEF D: LUFTFAHRZEUG E: AIRCRAFT
TS DIRIGEABLE D: LUFTSCHIFF E: AIRSHIPS

AVIATION R420
D: AVIATIK
E: AVIATION
TS VOL À VOILE D: GLEITFLUG E: GLIDING
VA AÉRONEF PLUS LOURD QUE L'AIR
D: LUFTFAHRZEUG SCHWERER ALS LUFT
E: HEAVIER-THAN-AIR AIRCRAFT

AVION T400
D: MOTORFLUGZEUG
E: AEROPLANES
NE aéronef plus lourd que l'air, propulsé, à aile fixe
TG AÉRONEF À AILE FIXE D: FLUGZEUG
E: FIXED-WING AIRCRAFT
TS AVION À RÉACTION D: DÜSENFLUGZEUG
E: JET AEROPLANES
AVION CARGO D: FRACHTFLUGZEUG
E: FREIGHT AEROPLANES

AVION À RÉACTION T400
D: DÜSENFLUGZEUG
E: JET AEROPLANES
TG AVION D: MOTORFLUGZEUG
E: AEROPLANES

AVION CARGO T400
D: FRACHTFLUGZEUG
E: FREIGHT AEROPLANES
TG AVION D: MOTORFLUGZEUG
E: AEROPLANES
VA FRET D: FRACHT E: FREIGHT

Cargaison EM FRET

DIRIGEABLE T1400
D: LUFTSCHIFF
E: AIRSHIPS
TG AÉROSTAT D: AEROSTAT
E: LIGHTER-THAN-AIR AIRCRAFT

FRET E440
D: FRACHT
E: FREIGHT
EP Cargaison
VA AVION CARGO D: FRACHTFLUGZEUG
E: FREIGHT AEROPLANES

HÉLICOPTÈRE T400
D: HUBSCHRAUBER
E: HELICOPTERS
TG AÉRONEF PLUS LOURD QUE L'AIR
D: LUFTFAHRZEUG SCHWERER ALS LUFT
E: HEAVIER-THAN-AIR AIRCRAFT

PLANEUR T400
D: SEGELFLUGZEUG
E: GLIDERS
TG AÉRONEF À AILE FIXE D: FLUGZEUG
E: FIXED-WING AIRCRAFT
TS VOL À VOILE D: GLEITFLUG E: GLIDING

VÉHICULE T400
D: FAHRZEUG
E: VEHICLES
TS AÉRONEF D: LUFTFAHRZEUG E: AIRCRAFT

VOL À VOILE R420
D: GLEITFLUG
E: GLIDING
TG AVIATION D: AVIATIK E: AVIATION
VA PLANEUR D: SEGELFLUGZEUG E: GLIDERS

Figure 5B(ii) — Alphabetical section (French as filing language)

AEROSTAT T600
E: LIGHTER-THAN-AIR AIRCRAFT
F: AÉROSTAT
OB LUFTFAHRZEUG E: AIRCRAFT F: AÉRONEF
UB LUFTSCHIFF E: AIRSHIPS F: DIRIGEABLE

AVIATIK R620
E: AVIATION
E: AVIATION
BF Flugwesen
UB GLEITFLUG E: GLIDING F: VOL À VOILE
VB LUFTFAHRZEUG SCHWERER ALS LUFT
E: HEAVIER-THAN-AIR AIRCRAFT
F: AÉRONEF PLUS LOURD QUE L'AIR

DÜSENFLUGZEUG T600
E: JET AEROPLANES
F: AVION À RÉACTION
OB MOTORFLUGZEUG E: AEROPLANES
F: AVION

FAHRZEUG T600
E: VEHICLES
F: VÉHICULE
UB LUFTFAHRZEUG E: AIRCRAFT F: AÉRONEF

Flugwesen BS AVIATIK

FLUGZEUG T600
E: FIXED-WING AIRCRAFT
F: AÉRONEF À AILE FIXE
OB LUFTFAHRZEUG SCHWERER ALS LUFT
E: HEAVIER-THAN-AIR AIRCRAFT
F: AÉRONEF PLUS LOURD QUE L'AIR
UB MOTORFLUGZEUG E: AEROPLANES
F: AVION
SEGELFLUGZEUG E: GLIDERS F: PLANEUR

FRACHT E640
E: FREIGHT
F: FRET
BF Kargo
Ladung
VB FRACHTFLUGZEUG E: FREIGHT
AEROPLANES F: AVION CARGO

FRACHTFLUGZEUG T600
E: FREIGHT AEROPLANES
F: AVION CARGO
OB MOTORFLUGZEUG E: AEROPLANES
F: AVION
VB FRACHT E: FREIGHT E: FRET

GLEITFLUG R620
E: GLIDING
F: VOL À VOILE
OB AVIATIK E: AVIATION F: AVIATION
VB SEGELFLUGZEUG E: GLIDERS F: PLANEUR

Gleitflugzeug BS SEGELFLUGZEUG

HUBSCHRAUBER T600
E: HELICOPTERS
F: HÉLICOPTÈRE
OB LUFTFAHRZEUG SCHWERER ALS LUFT
E: HEAVIER-THAN-AIR AIRCRAFT
F: AÉRONEF PLUS LOURD QUE L'AIR

Kargo BS FRACHT

Ladung BS FRACHT

LUFTFAHRZEUG T600
E: AIRCRAFT
F: AÉRONEF
OB FAHRZEUG E: VEHICLES F: VÉHICULE
UB AEROSTAT E: LIGHTER-THAN-AIR AIR-
CRAFT F: AÉROSTAT
LUFTFAHRZEUG SCHWERER ALS LUFT
E: HEAVIER-THAN-AIR AIRCRAFT
F: AÉRONEF PLUS LOURD QUE L'AIR
MILITÄRLUFTFAHRZEUG E: MILITARY AIR-
CRAFT E: AÉRONEF MILITAIRE

LUFTFAHRZEUG SCHWERER ALS LUFT T600
E: HEAVIER-THAN-AIR AIRCRAFT
F: AÉRONEF PLUS LOURD QUE L'AIR
OB LUFTFAHRZEUG E: AIRCRAFT F: AÉRONEF
UB FLUGZEUG E: FIXED-WING AIRCRAFT
F: AÉRONEF À AILE FIXE
HUBSCHRAUBER E: HELICOPTERS
F: HÉLICOPTÈRE
VB AVIATIK E: AVIATION F: AVIATION

LUFTSCHIFF T600
E: AIRSHIPS
F: DIRIGEABLE
OB AEROSTAT E: LIGHTER-THAN-AIR AIR-
CRAFT F: AÉROSTAT

MILITÄRLUFTFAHRZEUG T600
E: MILITARY AIRCRAFT
F: AÉRONEF MILITAIRE
OB LUFTFAHRZEUG E: AIRCRAFT F: AÉRONEF

MOTORFLUGZEUG T600
E: AEROPLANES
F: AVION
D nur für Luftfahrzeuge mit befestigten Flügeln, die schwerer als Luft sind
OB FLUGZEUG E: FIXED-WING AIRCRAFT
F: AÉRONEF À AILE FIXE
UB DÜSENFLUGZEUG E: JET AEROPLANES
F: AVION À RÉACTION
FRACHTFLUGZEUG E: FREIGHT AEROPLANES
F: AVION CARGO

SEGELFLUGZEUG T600
E: GLIDERS
F: PLANEUR
BE Gleitflugzeug
OB FLUGZEUG E: FIXED-WING AIRCRAFT
F: AÉRONEF À AILE FIXE
VB GLEITFLUG E: GLIDING F: VOL À VOILE

Figure 5C(ii) — Alphabetical section (German as filing language)

Figure 6A — Arrowgraph

Aerodynes USE HEAVIER-THAN-AIR AIRCRAFT

AEROPLANES T510
SN fixed-wing, powered, heavier-than-air aircraft
BT FIXED-WING AIRCRAFT
NT FREIGHT AEROPLANES
JET AEROPLANES

Aerostats USE LIGHTER-THAN-AIR AIRCRAFT

AIRCRAFT T510
BT VEHICLES
NT HEAVIER-THAN-AIR AIRCRAFT
LIGHTER-THAN-AIR AIRCRAFT
MILITARY AIRCRAFT

AIRSHIPS T510
BT LIGHTER-THAN-AIR AIRCRAFT

AVIATION R110
NT GLIDING
RT HEAVIER-THAN-AIR AIRCRAFT

Cargo USE FREIGHT

FIXED-WING AIRCRAFT T510
BT HEAVIER-THAN-AIR AIRCRAFT
NT AEROPLANES
GLIDERS

FREIGHT E116
UF Cargo
RT FREIGHT AEROPLANES

GLIDERS T510
BT FIXED-WING AIRCRAFT
RT GLIDING

GLIDING R110
BT AVIATION
RT GLIDERS

HEAVIER-THAN-AIR AIRCRAFT T510
UF Aerodynes
BT AIRCRAFT
NT FIXED-WING AIRCRAFT
HELICOPTERS
RT AVIATION

HELICOPTERS T510
BT HEAVIER-THAN-AIR AIRCRAFT

JET AEROPLANES T510
BT AEROPLANES

LIGHTER-THAN-AIRCRAFT T510
UF Aerostats
BT AIRCRAFT
NT AIRSHIPS

MILITARY AIRCRAFT T510
BT AIRCRAFT

VEHICLES T510
NT AIRCRAFT

Figure 6B — English alphabetical section

AEROSTAT T510
  OB LUFTFAHRZEUG
  UB LUFTSCHIFF

AVIATIK R110
  BF Flugwesen
  UB GLEITFLUG
  VB LUFTFAHRZEUG SCHWERER ALS LUFT

DÜSENFLUGZEUG T510
  OB MOTORFLUGZEUG

FAHRZEUG T510
  UB LUFTFAHRZEUG

Flugwesen BS AVIATIK

FLUGZEUG T510
  OB LUFTFAHRZEUG SCHWERER ALS LUFT
  UB MOTORFLUGZEUG
    SEGELFLUGZEUG

FRACHT E116
  BF Kargo
    Ladung
  VB FRACHTFLUGZEUG

FRACHTFLUGZEUG T510
  OB MOTORFLUGZEUG
  VB FRACHT

GLEITFLUG R110
  OB AVIATIK
  VB SEGELFLUGZEUG

Gleitflugzeug BS SEGELFLUGZEUG

HUBSCHRAUBER T510
  OB LUFTFAHRZEUG SCHWERER ALS LUFT

Kargo BS FRACHT

Ladung BS FRACHT

LUFTFAHRZEUG T510
  OB FAHRZEUG
  UB AEROSTAT
    LUFTFAHRZEUG SCHWERER ALS LUFT
    MILITÄRLUFTFAHRZEUG

LUFTFAHRZEUG SCHWERER ALS LUFT T510
  OB LUFTFAHRZEUG
  UB FLUGZEUG
    HUBSCHRAUBER
  VB AVIATIK

LUFTSCHIFF T510
  OB AEROSTAT

MOTORFLUGZEUG T510
  D nur für Luftfahrzeuge mit befestigten Flügeln, die schwerer als Luft sind
  OB FLUGZEUG
  UB DÜSENFLUGZEUG
    FRACHTFLUGZEUG

MILITÄRLUFTFAHRZEUG T510
  OB LUFTFAHRZEUG

SEGELFLUGZEUG T510
  BE Gleitflugzeug
  OB FLUGZEUG
  VB GLEITFLUG

Figure 6D — German alphabetical section

ELSEVIER

Information Processing and Management 40 (2004) 791–805

INFORMATION PROCESSING & MANAGEMENT

www.elsevier.com/locate/infoproman

Concept-based ranking: a case study in the juridical domain

Maria L. Silveira [a,b,*], Berthier Ribeiro-Neto [a,c,1]

[a] *Computer Science Department, Federal University of Minas Gerais, Brazil*
[b] *Prodabel—Information Technology Company of the City of Belo Horizonte, Minas Gerais, Brazil*
[c] *Akwan Information Technologies, Belo Horizonte, Minas Gerais, Brazil*

Available online 2 June 2004

Abstract

We explore the idea of automatically using the concepts of a thesaurus to improve search results. Our focus is on improving average precision figures, not recall. This is of interest because there is a tendency to accept the view that thesauri are recall-enhancing devices that are not good for improving precision figures.

In our approach, the query terms are used to match concepts in the thesaurus. These concepts are then used to find other related concepts (narrow, broad, synonym), which are interpreted as independent sources of evidential knowledge. Each source of evidence is used to produce a separate *concept-based ranking* of the documents in the collection. These partial rankings are then combined into a final ranking. For this, we use a Bayesian belief network.

To validate our ideas we do a case study in the juridical domain. Using a juridical thesaurus and a test collection containing more than 500 hundred thousand juridical documents, obtained from the legal court system in Brazil, we compare our concept-based ranking with the standard vectorial ranking. The results indicate improvements in average precision figures of roughly 30%.

© 2004 Elsevier Ltd. All rights reserved.

Keywords: Concept-based ranking; Bayesian networks; Juridical domain; Thesaurus

1. Introduction

Some 30 years ago, the idea of automating the process of searching for information in a document collection gained strength. Back then, the key notion revolved around the design of algorithms based on the searching processes used by specialists on bibliography. In this case, the documents of the collection were indexed using the concepts of a pre-defined controlled vocabulary. To search the collection the user selected concepts from the controlled vocabulary. The retrieval system then compiled the documents indexed by those concepts. As a result, automation consisted of creating inverted lists of concepts and of searching these inverted lists at query time.

* Corresponding author. Address: Prodabel—Information Technology Company of the City of Belo Horizonte, Minas Gerais, Brazil.

E-mail addresses: dilu@pbh.gov.br (M.L. Silveira), berthier@dcc.ufmg.br (B. Ribeiro-Neto).

URLs: http://www.dcc.ufmg.br, http://www.pbh.gov.br/prodabel, http://www.akwan.com.br.

[1] Partially supported by CNPq, Brazil, Individual Grant 300.188/95-1 and Grant 17.0435/01-6 (Project I^3DL).

0306-4573/$ - see front matter © 2004 Elsevier Ltd. All rights reserved.
doi:10.1016/j.ipm.2004.04.015

Since creating a controlled vocabulary of indexing concepts for a domain is a time consuming and expensive task, current Web searching mechanisms simply consider that all distinct words occurring in all pages are indexing terms. This simplifies the task but frequently leads to poor results.

In the last five years, the idea of applying concept-based indexing and searching to documents in the Web has emerged. The main drive has been towards creating specialized thesauri (i.e., hierarchical structures of concepts for a given area of knowledge), or borrowing a pre-existent thesaurus whenever it is available, and automating the process of indexing the documents with the concepts of the thesaurus. This process is normally referred to as categorization. If documents are categorized automatically, increased functionality is made available to the user such as interactivity, browsing, and selective querying.

If automatic categorization is not an option (because it does not work for the collection at hand or is expensive), one can still attempt to use the concepts of the thesaurus to improve search results. This is the problem we explore here. We investigate whether modifying the ranking with information provided by the concepts of a thesaurus, which we call *concept-based ranking*, leads to improved average precision. To validate our ideas, we experiment with a collection of juridical documents. In our experiments, the user does not need to be aware of the juridical thesaurus and its controlled vocabulary because the ranking modification is done automatically.

To combine information from the text of the documents with information from the thesaurus (its concepts, narrow concepts, broad concepts), we adopt the formalism of Bayesian belief networks. We propose a variant of the belief network model for information retrieval (IR) introduced in Ribeiro-Neto and Muntz (1996). The adoption of a Bayesian network model is important because it allows using probabilities, Bayesian inference, and a graphical representation of the key variables in the domain to model the IR problem. This is very useful to guide the process of gaining an intuitive understanding of the problem.

Previous results indicate that expanding the user query with the concepts of a thesaurus is effective to improve recall, but not necessarily precision (Kristensen, 1993). More recent results confirm that effective gains in recall can be generated, but report modest gains in precision (Greenberg, 2001a, 2001b). With specific collections, it is possible to obtain significant gains in average precision without improving precision at low recall. Our results here extend these findings by providing significant gains in average precision at all recall levels. Using a collection of queries and documents from the juridical domain, we observe gains in average precision around 30%. Our work extends the preliminary findings in Silveira, Ribeiro-Neto, Vale, and Assumpção (2003), in three directions, as follows.

First, the network model in Silveira et al. (2003) included just three sources of evidence: keywords, concepts from the juridical thesaurus, and narrow terms. In this paper, we discuss an extended network that includes six sources of evidence: keywords, concepts of the thesaurus, narrow terms, broad terms, synonyms, and related terms. This is important because it illustrates extensibility. Second, the results in Silveira et al. (2003) are preliminary. Our results here are much more extensive. We study the combination of keywords with five other sources of evidence. Dozens of experiments were run, before we reached the conclusions in this paper. Third, we show that the conventional procedure of automatic expansion of the original query (keywords), based on the other five sources of evidence, deteriorates retrieval performance for our juridical collection. This result confirms previous results in the literature. Following we show that our extended network model avoids this negative impact in retrieval performance. Our model allows combining the original query with the other sources of evidence available to yield gains in retrieval performance. This result is new and is not included in Silveira et al. (2003).

The main contributions of our work are as follows. First, we propose a vertical IR system for jurisprudence and evaluate its retrieval performance using a collection composed of hundreds of thousands of documents. We have no knowledge of other work that discusses experimentation with a large collection of juridical documents. Second, we show that average precision can be improved with the use of a specific-domain thesaurus, even at low recall levels. This is important because there are recent results, published in

2001, suggesting otherwise. Third, we show that a more structured approach, one that uses the formalism of Bayesian networks, has advantages over more ad hoc procedures such as the one of automatic query expansion. Our findings are validated with experiments based on a set of 30 queries obtained from a real Web site.

The article is organized as follows. In Section 2 we discuss related work. In Section 3 we describe the thesaurus we used in our work. In Section 4 we present our Bayesian network model for concept-based ranking. In Section 5 we present our experimental environment and discuss our experimental results. Our conclusions follow.

2. Related work

The importance of research in specialized (or, *vertical*) searching systems for the juridical area is growing, due to the increased availability of juridical document collections in the Web. To exemplify, Dempsey, Vreeland, Sumner, and Yang (2000) proposed a type of vertical system for searching juridical documents. In their system, the user can input three pieces of data describing his information need: (a) the type of document to be searched (juridical, legislative, etc.), the Court of his interest, and a query in natural language. This information is submitted to a pool of specialized search engines. The retrieved pages are indexed in a virtual collection that can be searched by the user. The authors argue that their strategy increases recall.

An interesting discussion on the role of taxonomies in legal information systems can be found in Bench-Capon and Visser (1997). The authors, who refer to taxonomies as explicit specifications of domain conceptualizations, discuss the advantages derived from producing a taxonomy for vertical systems, such as those for the legal domain.

On the commercial forefront, systems such as Westlaw and West Win have been designed to deal with collections of juridical documents. The two systems provide a search environment for juridical documents in authentic form (i.e., without headnotes, citations, and indexing terms). Additional material to the original document, when existent, is the result of an editorial effort. Not much is known about the architecture and internal features of these systems because documentation has not been published by West, as observed in Bing (1987). A third system that is little documented is the Lexis legal information system, as observed in Hafner (1987).

The authors in Greenleaf, Mowbray, and Tyree (1987) proposed a shell system suited to the development of legal expert systems, developed within the scope of the DataLex project. Their shell, which included functions specially tailored to support the requirements of legal expert systems, was used in a number of experimental application programs developed for teaching in Australian universities.

In Tong, Reid, Crowe, and Douglas (1987) the authors proposed a rule-based knowledge representation system called RUBRIC, which allows the user to stress a particular concept as well as to assign relative degrees of importance to distinct concepts. This is done by the specification of weights by the user. It is clearly a non-trivial query formulation process that requires training, patience, and good understanding of the set of concepts used to describe the application domain. No experimental data is provided.

In Daniels and Rissland (1997) the authors proposed to use knowledge derived from past cases in the SPIRE system to find documents or document passages associated with a problem situation. Case-based reasoning was used to generate the queries. The authors used a tiny database composed of just 20 documents to illustrate the operation of their system. Other study that considered application of case-based reasoning to the legal domain was presented in Mital, Stylianou, and Johnson (1991).

In Agosti, Colotti, and Gradenigo (1991) the authors proposed a legal retrieval system in which the documents are represented not only by terms, but also by structured data existing inside the documents, text fragments, and connections between documents, forming a hyperdocument. The system can also use

thesauri or a classification scheme to construct the hierarchy between concepts. The user can navigate in the structure existing in the documents and can create his own relationships between documents.

Many AI researchers proposed specialized searching systems for the juridical area (Bueno, von Wangenheim, Mattos, Hoeschl, & Barcia, 1999; Dick, 1987; Rose & Belew, 1989; Tong et al., 1987; Weber, 1999). Dick (1987) proposed a model to retrieve case law with emphasis on the development of a knowledge representation that is based on case grammar. Her focus is on finding arguments or part of arguments that could be useful to the searcher, but she did not show experimental results.

Rose and Belew (1989) proposed a combination of symbolic and connectionist AI techniques. Their argument is based on the observation that case law has both symbolic and subsymbolic information and that the use of these two types of information would lead to improved results. But they also did not show experimental results.

Bueno et al. (1999), Weber (1999) used case-based reasoning techniques to find similar cases in a juridical database. Their approach is based on the principle of analogy, which considers that similar problems have similar solutions. This reasoning is applied to past and new cases in the area of law. Weber (1999) illustrated her case with a very small collection of Brazilian juridical documents (just 3446 documents are considered) but also did not show experimental results. Bueno et al. (1999) proposed to use two types of controlled vocabulary. The first, named Controlled Vocabulary, contained a list of normative key-terms extracted from the specialized literature (Codes and Norms). The other, named Juridical Thesaurus, was constructed by law experts. Again, no experimental results were exhibited. An additional work that is not supported by experimental data is that of Tong et al. (1987).

In Smith (1997) the author discussed a search engine for information retrieval in vertical markets such as law and medicine—the Flexicon system. Flexicon uses a set of eight "lexicons" to assist the user with forming his query and browsing the database for cases that are related to his information need. However, experimental data and validation are simplistic. They discussed examples based on a single user information need in the paper.

Thesauri are domain specific hierarchical graph structures whose nodes are the key concepts of a specific area of knowledge. A thesaurus describes the main concepts of the area of interest and their relationships, which can be of three basic distinct types: equivalence, hierarchy, and associativity. In equivalence relationships, the concepts are synonyms, quasi-synonyms, and partial synonyms. In hierarchy relationships, the concepts have a subordination relationship, like broad and narrow terms. In associativity relationships, the concepts have a horizontal relationship, different from the previous ones, defined by specialists in the knowledge domain (Lancaster, 1991).

The thesauri were originally designed to help librarians and specialists with the task of indexing the documents of a collection. Nowadays, thesauri are also being used to help users search for information of interest. While searching, the users can explore the concepts of the thesaurus and browse their relationships. This might lead the user to new unforeseen concepts related to his original query that might be useful in the query formulation process (Gilchrist, 1971; Shiri & Revie, 2000).

The traditional process of constructing a thesaurus is slow and laborious, since it is usually hand-crafted by specialists on the knowledge domain. An alternative is to attempt to construct the thesaurus automatically, a topic that we do not discuss here. Some references for this are Chen, Yim, Fye, and Schatz (1995), Qiu and Frei (1993), Crouch and Yang (1992), and Crouch (1988).

Many authors used a thesaurus to expand the user query automatically (Bing, 1987; Greenberg, 2001a, 2001b; Kristensen, 1993; Mandala, Tokunaga, & Tanaka, 1999, 2000). In this case, concepts of the thesaurus that are related to the user query are factored in and processed together with the original query terms.

Greenberg (2001a, 2001b) investigated the use of a business thesaurus for query expansion. The experiments were executed over a collection of business related documents. The queries were real queries selected from questions formulated by business students. The author considered only queries that could be

mapped to a concept in the thesaurus. She concluded that, if the focus is on improving precision figures, synonyms and narrow terms should be used for query expansion. Otherwise, if the focus is on improving recall figures, then related terms and broad terms should be used for query expansion.

Kristensen (1993) studied the use of synonym, narrow and related terms in query expansion in a full text newspaper database. She found enhancements in recall and reduction in precision, and concluded that a thesaurus is "clearly a recall-enhancing tool".

Mandala et al. (1999, 2000) used three types of thesauri to select terms for automatic query expansion. The types of thesauri were hand-crafted, generated automatically using co-occurrence of terms and generated automatically using predicate-argument logic to gather linguistic relations. The authors developed algorithms to estimate the probability that the terms of each thesaurus are good candidates for query expansion. They showed that the combination of the three types of thesauri yields better results.

Bing (1987) proposed a norm-based thesaurus to be used in conceptual searching. This thesaurus has a hierarchical structure that "reflects the relation between the statutes and the valid legal norms". The key problem is to reduce the natural language of the statutes to an structure that is sufficiently formalized to permit conceptual searching. The paper includes no experimental results.

Clark, Thompson, Holmback, and Duncan (2000) used a thesaurus to find knowledge expert persons in a large company. In their system, the knowledge expert person used the thesaurus to describe his expertise profile. The description is recorded like a document. The search feature is used to find persons that match a desired profile described using concepts of the thesaurus.

Since the framework of Bayesian networks is used as a theoretical foundation in our work, we also review related work in this area. Bayesian networks have been used extensively to model IR related problems since the 1990s. They have been applied to query expansion (Turtle & Croft, 1990), logs of past queries (Ribeiro-Neto & Muntz, 1996; Ribeiro-Neto, Silva, & Muntz, 2000), structured querying (Croft, Turtle, & Lewis, 1991), modeling of the link structure of the Web (Calado, Ribeiro-Neto, Ziviani, Moura, & Silva, 2001; Silva, Ribeiro-Neto, Calado, Moura, & Ziviani, 2000), relevance feedback (Haines & Croft, 1993), modeling of topics related to the user query (Turtle & Croft, 1991), retrieval of medical documents based on thesauri (Vale, Ribeiro-Neto, Lima, Laender, & Freitas-Junior, 2003). In all these studies, the framework of Bayesian networks has been shown valuable. The models presented have led to gains in retrieval performance and have allowed a more intuitive understanding of the problem under consideration.

The work of Campos, Fernández-Luna, and Huete (1998) proposed a method to construct a Bayesian network that represents some of the term relationships in a document collection. This Bayesian network is then used as a thesaurus. The authors report experiments with three document collections: ADI, Crafield, and MedLars. For two of these collections, they report significant gains in average precision figures of roughly 16% and 13%. For the third collection, the gains are modest. The authors concluded that the improvements are collection dependent.

The authors in Campos, Fernández-Luna, and Huete (2003) proposed a Bayesian network to model dependencies among index terms, referred to as BNRM—Bayesian network retrieval model. They modeled the query as new evidence to be introduced in the system. As in Campos et al. (1998), their objective was to construct a collection dependent thesaurus using Bayesian networks. The authors tested their model with five reference collections. For three of them they observed gains in precision figures and concluded that the results are collection dependent.

In Rajashekar and Croft (1995) the authors made a series of experiments involving thesaurus and Bayesian inference networks. These experiments included stemmed text, keywords, Boolean and weighted queries, all combined with thesaurus queries in an inference network. They concluded that the combination of multiple evidences, both in the queries and in the index, leads to superior results. This is also observed in our results in this paper. They have also concluded that the use of a thesaurus was not effective for improving results. Our conjecture to explain their findings is that they used a generic thesaurus with a generic collection. In this case, the inherent polysemy of generic nouns introduces noise that leads to poor

retrieval performance. Our results here complement their findings. We show that a specific-domain thesaurus applied to a collection of documents from the same area of knowledge, such as the juridical area, might lead to good gains in average precision figures.

In here, we extend the preliminary work of Silveira et al. (2003) in three directions. First, while the network model in Silveira et al. (2003) included just three sources of evidence (keywords, concepts from the juridical thesaurus, and narrow terms), the network model in this work includes six sources of evidence (keywords, concepts of the thesaurus, narrow terms, broad terms, synonyms, and related terms), showing that our approach can be naturally extended to include additional sources of evidence. Second, our results here are much more extensive. We study the combination of keywords with the other five sources of evidence, performing dozens of experiments to reach our conclusions. Third, we show that the conventional procedure of automatic expansion of the original query (keywords) deteriorates retrieval performance for our juridical collection, while our extended network model avoids this negative impact in retrieval performance. Our model allows combining the original query with the other sources of evidence available to yield gains in retrieval performance. This result is new and is not included in Silveira et al. (2003).

Our focus here is on investigating whether evidential information obtained from a thesaurus can be used to improve the precision of search results. Contrary to the work in Greenberg (2001a, 2001b), our approach uses a model of the IR system based on Bayesian networks. The network is used to combine the original query with various sources of evidence obtained from a thesaurus (concepts related to the query, narrow concepts, broad concepts). The pieces of evidence are then summed up by a single ranking formula. Without such a formalism, working with just a crude vectorial representation of the problem, for instance, we would have to rely on a non-intuitive procedure of manipulating a set of term weights somehow, in an attempt to generate better results.

3. The CJF thesaurus

To corroborate our claims, we do a case study in the juridical area. We use a manually constructed juridical thesaurus to find concepts related to the original query. The idea is that these concepts can be used to improve search results. We adopt the thesaurus designed by the Federal Justice Council in Brazil, which we now discuss.

The Brazilian Federal Justice Council (CJF) is a regulatory institution that supervises the system of Federal Courts in Brazil and their operations (CJF, 2002). Since 1993, the CJF has coordinated the design, specification, construction, and maintenance of a thesaurus for Brazilian court rulings, which we refer to as the CJF thesaurus. The motivation was to construct a tool to control the vocabulary used by indexing specialists when constructing the Indexing Section [2] of the juridical document. The CJF thesaurus comprises many fields of Law, as for example, Criminal Law, Civil Law, Public Law, Commercial Law, Administrative Law, Constitutional Law, International Law, among others.

The construction process occurred between 1993 and 1997, with the coordination of the CJF and the participation of specialists from the Federal Justice system, the five Brazilian Federal Courts of Appeals (TRFs). The technical people working on the CJF thesaurus had a background on the areas of information science and law. Their work followed a *bottom up* approach (Lancaster, 1986), which consisted of surveying the juridical literature [3] to find the concepts considered to be useful and meaningful for indexing and retrieval. This approach is known as the *literary warrant* (Lancaster, 1986) (compare to the *user warrant* in

[2] After judgment, the juridical document is enriched with a section called Indexing Section created with the purpose to improve retrieval.

[3] Particularly Codes, i.e., collections of laws addressing a matter that is object of a field of Law, and Doctrines, i.e., principles that allow interpreting the juridical science.

which the concepts are obtained by surveying the users' vocabulary). Each field of Law was attributed to a participating group, which worked in separate and built a mini-thesaurus for that field. Following, the groups met to discuss the union of their mini-thesaurus and to define a common labeling procedure for the concepts and the relationships that compose the CJF thesaurus. A maintenance process was established. Proposals for updating the CJF thesaurus are submitted to a coordinating group that votes on the acceptance or rejection of the proposed updates.

The CJF thesaurus has 8357 juridical concepts organized hierarchically. From these concepts, 6103 are classified as narrow terms (NT), 891 as broad terms (BT), 7301 as related terms (RT), and 1702 as synonyms. Among the synonyms, 674 are classified as preferred (descriptors) and 1208 as non-preferred (non-descriptors). We note that a concept classified as a narrow term of a concept may be classified as broad term of another and as a related term of a third one.

4. A belief network for concept-based ranking

Our approach is to combine distinct sources of evidential information related to the user query, such as synonyms, narrow concepts, and broad concepts, and use this evidence to improve query results. For this, we adopt a modeling framework based on Bayesian belief networks.

Bayesian networks are useful because they allow explicitly representing the independencies among the random variables in a domain, which is critical to allow efficient processing and more precise results. Further, quantitative metrics such as partial rankings and term weights are modeled as probability functions, a procedure that follows well established operations and is well understood. We can say that Bayesian networks provide a framework that allows representing all the key variables that interfere with the retrieval results of an IR system. That is, it provides a formal underpinning to the IR problem.

In the belief network model (Ribeiro-Neto & Muntz, 1996), which we adopt in this work, the probability ranking function is represented through a directed acyclic graph whose nodes model the variables of the problem. Relationships among these variables are modeled as directed edges that represent causal dependencies among the linked variables. The strength of these dependencies are expressed by conditional probabilities.

In the network of Fig. 1, each node D_j models a document d_j, the node Q models the user query q, and the K_i nodes model the keywords in the collection. The set K is used to refer to any of the possible states of the set of *root* nodes K_i. A *binary* random variable is associated with the node Q, which is also denoted by Q. [4] The variable Q is 1, denoted by q, to indicate that Q is *on* and Q is 0, denoted by $\bar{q}$, to indicate that Q is *off*. These two states correspond to confirming/observing (or not confirming/observing) the query, given that a set of keywords has been pre-selected. Analogously, a *binary* random variable D_j is associated with the document node D_j. The variable D_j is 1, denoted by d_j, to indicate that D_j is *on* and D_j is 0, denoted by $\bar{d}_j$, to indicate that the variable D_j is *off*. As in the case of the query, these two states indicate that the document is observed (or not observed), given a subset of pre-selected keywords. A *binary* random variable K_i is also associated with each keyword K_i. All of these variables are binary because this is simple and provides enough semantics for modeling the information retrieval problem. Varying degrees of relevance are represented in the model as conditional probabilities, as we discuss below.

In Fig. 1, the keywords (KY) that compose the user query Q are mapped directly into thesaurus concepts (CC). This is done automatically by matching query terms to the concepts in the thesaurus. The matching concepts form the concept-based version of the query, which is represented in the network by the set of

[4] In this notation it should always be clear whether we are referring to the query, to the node in the network, or to its associated binary variable.

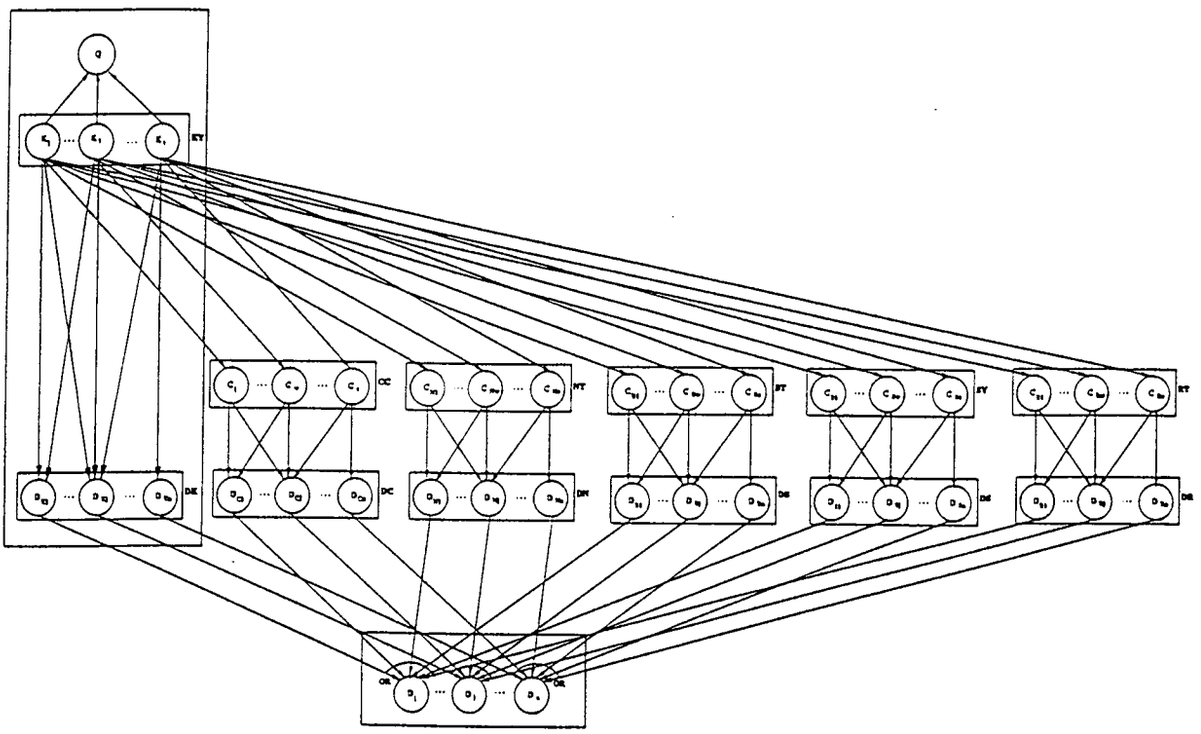


Fig. 1. Bayesian belief network for concept-based ranking using the CJF thesaurus.

concepts C. Each of the concepts in this set, on its turn, can be mapped into synonyms (SY), related terms (RT), narrow terms (NT), or broad terms (BT). This is also done automatically by following the relationships in the thesaurus. Such mappings are represented in the network by the sets of evidence S, R, N, and B, respectively. To simplify the representation, we model these evidences as sets of independent variables. This allows fast computation while capturing enough semantics to allow improving search results.

The nodes D_{Kj}, D_{Cj}, D_{Sj}, D_{Rj}, D_{Nj}, and D_{Bj} represent the document D_j in distinct contexts. The node D_{Kj} is used to represent the document D_j in the context of a keyword-based retrieval process (i.e., the user query is considered as composed of keywords and is processed using the standard cosine formula of the vector model). The node D_{Cj} is used to represent the document D_j in the context of a concept-based retrieval process. If there is a link from a concept C_i to a node D_{Cj}, then the text of the document D_j contains a match to the concept C_i. We modeled D_{Cj} and D_{Kj} separately in our network to allow evaluating the impact of concept-based retrieval (versus keyword-based retrieval) in the quality of the results. The node D_{Nj} is used to represent the document D_j in the context of a retrieval process based on narrow concepts (which are related to the original user query). As before, there is a link from a narrow concept C_{Ni} to a node D_{Nj} if the text of the document D_j contains a match to the narrow concept C_{Ni}. The nodes D_{Rj}, D_{Bj}, and D_{Sj} are used to represent the document D_j when related, broad and synonym concepts, respectively, are used as queries.

The documents on the set DK are ranked according to the standard vector model (i.e., the cosine formula applied to keywords). The documents in the sets DC, DS, DR, DN, and DB are ranked using the cosine formula applied to concepts. These sets of ranked documents represent additional evidence that can be accumulated to yield a new ranking (which, we presume, has higher precision).

In our belief network, the rank of a document D_j is computed as follows:

$$P(d_j \mid q) = \eta \sum_{\forall k \in \mathbf{K}} P(d_j \mid k) P(q \mid k) P(k) \tag{1}$$

Assuming that the unique keywords of interest are the query keywords, we define $P(q \mid k)$ as:

$$P(q \mid k) = \begin{cases} 1 & \text{if } \mathbf{K} = k_q \\ 0 & \text{otherwise} \end{cases} \tag{2}$$

where k_q is a state of the keywords in $\mathbf{K}$ in which the unique active nodes are those present in the query. This allows rewriting Eq. (1) as:

$$P(d_j \mid q) \sim P(d_j \mid k_q) P(q \mid k_q) P(k_q) \tag{3}$$

where we deleted the normalizing constant η, because it has no effect on the ranking.

We observe that $P(d_j \mid k_q)$ depends on the evidence obtained from the different concepts and their relationships in the thesaurus. These evidences are used to enrich the network with distinct representations of the original query. For each such representation, a ranking of documents is generated (the corresponding answer sets are D_{Cj}, D_{Sj}, D_{Rj}, D_{Nj}, and D_{Bj}). These rankings are interpreted as sources of evidence on the final relevance of the documents to the original query. To combine these sources of evidence we use a disjunctive operator because it yields better results (Ribeiro-Neto & Muntz, 1996; Ribeiro-Neto et al., 2000; Silva et al., 2000). Thus, the document node D_j accumulates all the ranking evidence through a disjunction of the beliefs associated with the nodes D_{Kj}, D_{Cj}, D_{Sj}, D_{Rj}, D_{Nj}, and D_{Bj}. This allows rewriting Eq. (3) as:

$$\begin{aligned} P(d_j \mid q) \sim [1 - (1 - P(dk_j \mid k_q)) \times (1 - P(dc_j \mid c_q)) \times (1 - P(dn_j \mid n_q)) \times (1 - P(db_j \mid b_q)) \\ \times (1 - P(ds_j \mid s_q)) \times (1 - P(dr_j \mid r_q))] \times P(q \mid k_q) \times P(k_q) \end{aligned} \tag{4}$$

where c_q, n_q, b_q, s_q, and r_q represent states of the sets of random variables C, N, B, S, and R, respectively, in which the only active nodes are those associated directly or indirectly with the keywords in the original user query. We notice that no specific weights are assigned to any of the components of our ranking equation. Given prior knowledge on the application domain, one might attempt to weigh more the more significative components of our equation, an issue that we do not discuss here.

To compute conditional probabilities we proceed as follows. Consider, for instance, the conditional probability $P(dk_j \mid k_q)$. We have

$$P(dk_j \mid k_q) = \frac{\sum_{i=1}^{t} w_{ij} \times w_{ik}}{\sqrt{\sum_{i=1}^{t} w_{ij}^2} \times \sqrt{\sum_{i=1}^{t} w_{ik}^2}} \tag{5}$$

where w_{ik} and w_{ij} are *tf-idf* weights (Salton & McGill, 1983) for the query q and the document d_j, respectively. Eq. (5) computes a ranking as the standard vector model. The other probabilities are computed analogously, but considering weights relative to the concepts, narrow concepts, broad concepts, etc., of the CJF thesaurus.

Eq. (5) is the ranking formula of our Bayesian model for a juridical digital library. Once the conditional probabilities have been computed, the ranking computation is done in linear time. In practice, this means that the time to compute our concept-based ranking is a linear function of the time to compute a standard vectorial ranking.

Our ranking formula allows combining evidence in several ways. For instance, consider, for a moment, that we are interested only in the results yielded by the vector model. To obtain this effect, we define:

$$P(dc_j \mid k_{cq}) = 0; \quad P(dn_j \mid k_{nq}) = 0; \quad P(db_j \mid k_{bq}) = 0; \quad P(ds_j \mid k_{sq}) = 0; \quad P(dr_j \mid k_{rq}) = 0$$

As a result, the ranking $P(d_j \mid q)$ becomes

$$P(d_j \mid q) \sim P(dk_j \mid k_q) \times P(q \mid k_q) \times P(k_q)$$

which computes a vectorial ranking. To consider the combination of keyword-based and concept-based retrieval, we define:

$$P(dn_j \mid k_{nq}) = 0; \quad P(db_j \mid k_{bq}) = 0; \quad P(ds_j \mid k_{sq}) = 0; \quad P(dr_j \mid k_{rq}) = 0$$

As a result,

$$P(d_j \mid q) \sim [1 - (1 - P(dk_j \mid k_q)) \times (1 - P(dc_j \mid k_{cq}))] \times P(q \mid k_q) \times P(k_q)$$

which yields a ranking that combines keyword-based and concept-based retrieval.

Further, any combination of evidences 2×2, 3×3, and so on, can be evaluated by properly defining the related conditional probabilities. Also, we observe that Eq. (5) allows us to compute 63 distinct ranking variations.

5. Experimental results

5.1. The juridical reference collection

Our evaluation is carried out using a reference collection composed of Brazilian juridical documents, of a set of real queries, and of a set of relevant documents for each of them, determined by a Brazilian lawyer. The collection is composed of 552,573 juridical documents from the Supreme Federal Court (STF) (STF, 2002), the Superior Court of Justice (STJ) (STJ, 2002), and the five Federal Courts of Appeal (TRF1, 2002; TRF2, 2002; TRF3, 2002; TRF4, 2002; TRF5, 2002). The average document size is 164.68 words.

The test queries were selected from a set of real requests made by users of the CJF Internet site (CJF, 2002). The queries selected, in number of 30, satisfied the following conditions: all queries contained at least one term that could be mapped into concepts of the CJF thesaurus, and the mapped concepts included all relationships i.e., synonyms, narrow terms, broad terms, and related terms. [5] The terms in the set **K** were obtained by eliminating stopwords from the query text. The concepts in the set **C** were obtained by directly matching the query terms to concept labels of the CJF thesaurus. The concept sets referred to as **S**, **N**, **B**, and **R** were obtained from the set **C** by following the thesaurus relationships. As a result of this mapping procedure, each test query q was represented in our belief network model by six distinct subqueries: K_q, C_q, N_q, B_q, S_q, and R_q. The average size of each of these subqueries was as follows: 9.03 keywords (KY), 6.40 concepts (CC), 3.46 synonyms (SY), 29.96 narrow terms (NT), 2.16 broad terms (BT), and 40.50 related terms (RT).

The evaluation of the relevant documents for each test query was done using the well known *pooling* method (Voorhees & Harman, 1996). The pool of candidate documents was composed of the top-50 ranked documents retrieved by each ranking variation generated by our belief network model. The pool was evaluated by a lawyer, who classified each document as relevant or non-relevant to the query. Since he did not know which search strategy retrieved a given document, his evaluation was necessarily non-biased.

[5] These decisions were made in order to verify the thesaurus contribution.

5.2. Experimental analysis

Our experiments were done in two phases. In the first phase, we explored the impact in search results of our concept-based ranking approach. This impact was quantified in terms of average precision figures relative to the vector model, which we take as a baseline. In the second phase, we compare the results of our concept-based ranking strategy with the more common procedure of automatic query expansion (in which concepts of the CJF Thesaurus are used to expand the original query).

5.2.1. Experiments using each source of evidence in separate

Fig. 2 illustrates the results generated by each source of evidence in separate, i.e., keywords (KY), mapped concepts (CC), and associated relationships (SY, NT, BT, RT). As we see, CC is an interesting source of evidence that yields results that are superior to those provided by KY. The relationships (SY, NT, BT, and RT) are not good sources of evidence for our test collection. They are too noisy retrieving an excessive number of new documents that are not of interest.

5.2.2. Experiments combining the sources of evidence

In Fig. 3 we show the results of the Bayesian combination of keywords with their associated concepts and relationships. We see that KY+CC provides a combination of evidence that leads to improved results. In terms of average precision, it is 31% superior to KY and 12% superior to CC. We argue that this gain is obtained from the combination of user provided knowledge (KY) with expert knowledge (CC) encoded in the thesaurus. The combinations of the original query with narrow terms (KY+NT) and with related terms (KY+RT) preserve the results obtained by the use of KY in isolation (vector model). The combinations with synonyms (KY+SY) and with broad terms (KY+BT) cause loss in precision.

We also experimented with combinations of evidence 3×3, 4×4, etc., in our network. Whenever a combination included the evidences KY and CC, average precision figures were superior to those provided by the use of KY only. However, none of these higher order combinations yielded better results than the combination KY+CC. Thus, this combination represents our best results.

We observe that the lack of improvements associated with four of our combinations (KY+NT, KY+RT, KY+SY, and KY+BT) is due to the low average precision figures provided by each of the respective sources

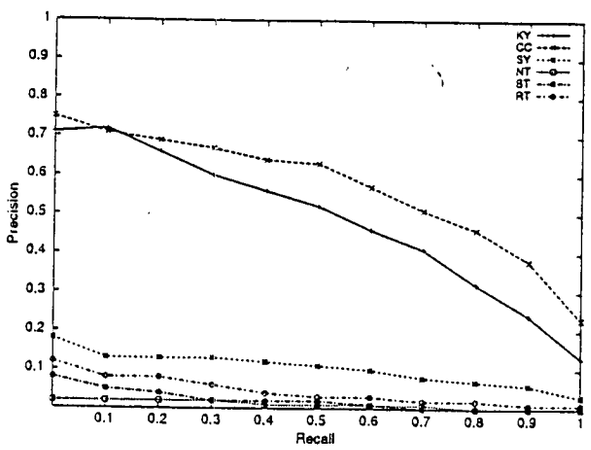


Fig. 2. Results generated by each source of evidence in separate, i.e., keywords (KY), concepts (CC), and relationships (SY, NT, BT, RT).

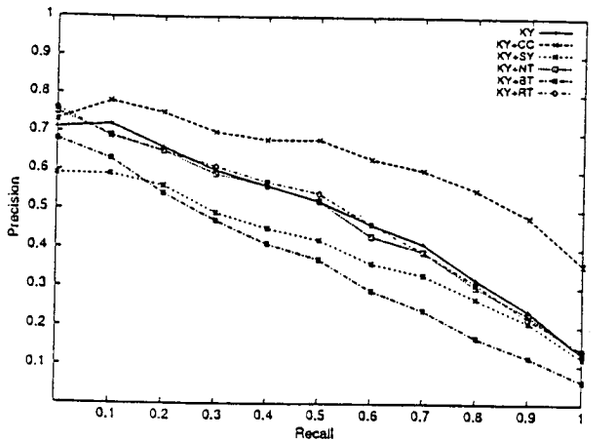


Fig. 3. Results generated by combining keywords with other sources of evidence, 2×2.

of evidence, not to our Bayesian framework. In fact, when a source of evidence is representative of the user interest (i.e., when it yields average precision figures that approximate the results provided by our baseline), its inclusion in our Bayesian network leads to improved results, as observed in the case for CC (related concepts).

5.2.3. Comparing Bayesian combination with automatic query expansion

To illustrate the difference between our concept-based ranking approach and the more common procedure of automatic query expansion, we compare results. In concept-based ranking, we combine the search results generated by various subqueries (one for each source of evidence considered). In automatic query expansion, we expand the original query with related concepts (and their relationships) and process the expanded query. Fig. 4 shows the results generated by expanding the original query with related concepts and their relationships and by combining all sources of evidence in our Bayesian network. While the

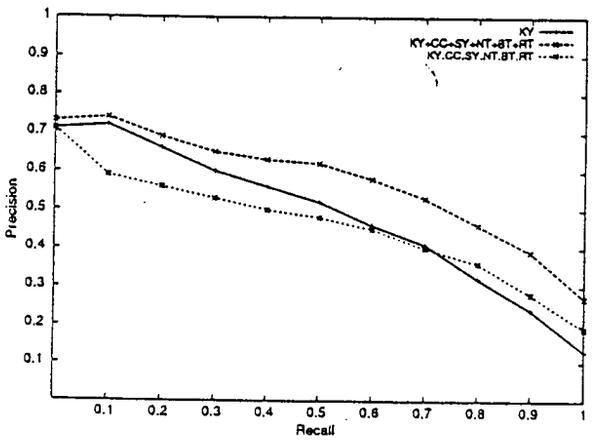


Fig. 4. Results using all sources of evidence: (a) combining them in our network, labelled as KY+CC+SY+NT+BT+RT, and (b) automatically expanding the original query with all sources of evidence, labelled as KY.CC.SY.NT.BT.RT.

combination of all sources of evidence leads to worst results than those provided by the combination KY+CC, [6] we observe that the procedure of automatic query expansion leads to a loss in average precision figures.

These late results corroborate previous results in the literature (Greenberg, 2001a, 2001b; Kristensen, 1993) that show that the indiscriminate usage of thesaurus evidences for query expansion deteriorates the quality of the answer set. This is because the addition of new terms to the original query tends to be a noisy procedure, since it occurs at the syntactic level of words. Our Bayesian network, on the other hand, operates at the higher level of ranked subsets of documents and yields better results.

6. Conclusions

In this work we explored the idea of automatically using a thesaurus to improve search results. In our approach, the query terms are used to match concepts in the thesaurus. These concepts are then used to find other related concepts (narrow, broad, synonym) in the thesaurus. Each set of related concepts is interpreted as an independent source of evidential knowledge. Independent sources of evidence are treated as subqueries and processed separately, generating partial answer sets. These are then combined in a Bayesian belief network to yield a final ranking, which we call *concept-based ranking*.

To validate our ideas we did a case study in the juridical domain. Using a juridical thesaurus and a test collection containing more than 500 hundred thousand juridical documents, obtained from the legal court system in Brazil, we compared our concept-based ranking with the standard vectorial ranking. We found improvements in average precision figures of roughly 30%. These results are new and pose a challenge to a somewhat consolidated view that thesauri are good for enhancing recall figures, but not for improving precision.

References

Agosti, M., Colotti, R., & Gradenigo, G. (1991). A two-level hypertext retrieval model for legal data. In *Proceedings of the 14th annual international ACM SIGIR conference on research and development in information retrieval* (pp. 316–325). ACM Press.

Bench-Capon, T., & Visser, P. (1997). Ontologies in legal information systems: The need for explicit specifications of domain conceptualisations. In *Proceedings of the sixth international conference on artificial intelligence and law* (pp. 132–141). ACM Press.

Bing, J. (1987). Designing text retrieval systems for conceptual searching. In *Proceedings of the first international conference on artificial intelligence and law* (pp. 43–51). ACM Press.

Bueno, T. C. D., von Wangenheim, C. G., Mattos, E. S., Hoeschl, H. C., & Barcia, R. M. (1999). JurisConsulto: Retrieval in jurisprudencial text bases using juridical terminology. In *Proceedings of the seventh international conference on artificial intelligence and law* (pp. 147–155). ACM Press.

Calado, P., Ribeiro-Neto, B., Ziviani, N., Moura, E., & Silva, I. (2001). Local versus global link information in the web. *ACM Transactions on Information Systems, 21*(1), 42–63.

Campos, L. M., Fernández-Luna, J. M., & Huete, J. F. (1998). Query expansion in information retrieval systems using a Bayesian network-based thesaurus. In *Proceedings of the 14th uncertainty in artificial intelligence conference* (pp. 53–60).

Campos, L. M., Fernández-Luna, J. M., & Huete, J. F. (2003). The BNR model: Foundations and performance of a Bayesian network-based retrieval model. *International Journal of Approximate Reasoning, 34*(2–3), 265–286.

Chen, H., Yim, T., Fye, D., & Schatz, B. (1995). Automatic thesaurus generation for an electronic community system. *Journal of the American Society for Information Science, 46*(3), 175–193.

CJF (2002). Federal Justice Council of the Federative Republic of Brazil. Available: www.cjf.gov.br, 2002.

Clark, P., Thompson, J., Holmback, H., & Duncan, L. (2000). Exploiting a thesaurus-based semantic net for knowledge-based search. In *Proceedings of the 12th conference on innovative applications of AI (AAAI/IAAI00)* (pp. 988–995).

[6] Because there are sources of evidence that are weak and contribute negatively in the summation.

TRF1 (2002). Federal Court of Appeals of the First Region of the Federative Republic of Brazil. Available: www.trf1.gov.br, 2002.
TRF2 (2002). Federal Court of Appeals of the Second Region of the Federative Republic of Brazil. Available: www.trf2.gov.br, 2002.
TRF3 (2002). Federal Court of Appeals of the Third Region of the Federative Republic of Brazil. Available: www.trf3.gov.br, 2002.
TRF4 (2002). Federal Court of Appeals of the Fourth Region of the Federative Republic of Brazil. Available: www.trf4.gov.br, 2002.
TRF5 (2002). Federal Court of Appeals of the Fifth Region of the Federative Republic of Brazil. Available: www.trf5.gov.br, 2002.
Turtle, H. R., & Croft, W. B. (1990). Inference networks for document retrieval. In *Proceedings of the thirteenth international ACM SIGIR conference on research and development in information retrieval, Brussels, Belgium* (pp. 1–24).
Turtle, H. R., & Croft, W. B. (1991). Evaluation of an inference network-based retrieval model. *Transaction on Information Systems, 9*(3), 187–222.
Vale, R. F., Ribeiro-Neto, B., Lima, L., Laender, A., & Freitas-Junior, H. (2003). Improving text retrieval in medical collections through automatic categorization. In *Proceedings of the string processing and information retrieval symposium (SPIRE)* (pp. 197–210).
Voorhees, E. M., & Harman, D. (1996). Overview of the fifth text retrieval conference. In *Proceedings of the fifth text retrieval conference (TREC-5)* (pp. 1–28). Gaithersburg, MD 20899: National Institute of Standards and Technology.
Weber, R. (1999). Intelligent jurisprudence research: A new concept. In *Proceedings of the seventh international conference on artificial intelligence and law* (pp. 164–172). ACM Press.

Croft, W. B., Turtle, H. R., & Lewis, D. D. (1991). The use of phrases and structured queries in information retrieval. In *Proceedings of the 14th annual international ACM SIGIR conference on research and development in information retrieval* (pp. 32–45). ACM Press.

Crouch, C. J. (1988). A cluster-based approach to thesaurus construction. In *Eleventh international conference on research and development in information retrieval* (pp. 309–320).

Crouch, C. J., & Yang, B. (1992). Experiments in automatic statistical thesaurus construction. In *15th annual international ACM SIGIR conference* (pp. 77–88).

Daniels, J. J., & Rissland, E. L. (1997). What you saw is what you want: Using cases to seed information retrieval. In *Proceedings of the second international conference on case-based reasoning (ICCBR-97), Providence, RI, USA.*

Dempsey, B. J., Vreeland, R. C., Sumner, R. G., Jr., & Yang, K. (2000). Design and empirical evaluation of searh software for legal professionals on the WWW. *Information Processing and Management, 36*, 253–273.

Dick, J. P. (1987). Conceptual retrieval and case law. In *Proceedings of the first international conference on artificial intelligence and law* (pp. 106–115). ACM Press.

Gilchrist, A. (1971). *The thesaurus in retrieval.* London: Aslib.

Greenberg, J. (2001a). Automatic query expansion via lexical–semantic relationships. *Journal of the American Society for Information Science and Technology, 52*(5), 402–415.

Greenberg, J. (2001b). Optimal query expansion (QE) processing methods with semantically encoded structured thesauri terminology. *Journal of the American Society for Information Science and Technology, 52*(6), 487–498.

Greenleaf, G., Mowbray, A., & Tyree, A. (1987). Expert systems in law: The datalex project. In *Proceedings of the first international conference on AI and law, Boston, USA* (pp. 9–17).

Hafner, C. (1987). Conceptual organization of case law knowledge bases. In *Proceedings of the first international conference on AI and law, Boston, USA* (pp. 35–42).

Haines, D., & Croft, W. B. (1993). Relevance feedback and inference networks. In *Proceedings of the sixteenth annual international ACM/SIGIR conference on research and development in information retrieval, Pittsburgh, PA, USA* (pp. 2–11).

Kristensen, J. (1993). Expanding end-users' query statements for free text searching with a search-aid thesaurus. *Information Processing & Management, 29*(6), 733–744.

Lancaster, F. W. (1986). *Vocabulary control for information retrieval.* Arlington, VA: Information Resources Press.

Lancaster, F. W. (1991). *Indexing and abstracting in theory and practice.* London: Library Association Publishing.

Mandala, R., Tokunaga, T., & Tanaka, H. (1999). Combining multiple evidence from different types of thesaurus for query expansion. In *Proceedings of the 22nd annual international ACM/SIGIR conference on research and development in information retrieval* (pp. 191–197).

Mandala, R., Tokunaga, T., & Tanaka, H. (2000). Query expansion using heterogeneous thesauri. *Information Processing & Management, 36*(2000), 361–378.

Mital, V., Stylianou, A., & Johnson, L. (1991). Conceptual information retrieval in litigation support systems. In *Proceedings of the third international conference on AI and law* (pp. 235–243). England: Oxford.

Qiu, Y., & Frei, H. P. (1993). Concept based query expansion. In *Proceedings of the sixteenth annual international ACM SIGIR conference on research and development in information retrieval* (pp. 160–169).

Rajashekar, T. B., & Croft, W. B. (1995). Combining automatic and manual index representations in probabilistic retrieval. *Journal of the American Society of Information Science, 46*(4), 272–283.

Ribeiro-Neto, B., & Muntz, R. (1996). A belief network model for IR. In *ACM conference on research and development in information retrieval—SIGIR96* (pp. 253–260).

Ribeiro-Neto, B., Silva, I., & Muntz, R. (2000). Bayesian network models for IR. In F. Crestani & G. Pasi (Eds.), *Soft computing in information retrieval techniques and applications* (pp. 259–291). Springer-Verlag.

Rose, D. E., & Belew, R. K. (1989). Legal information retrieval a hybrid approach. In *Proceedings of the second international on conference on artificial intelligence and law* (pp. 138–146). ACM Press.

Salton, G., & McGill, M. J. (1983). *Introduction to modern information retrieval.* New York: McGraw-Hill.

Shiri, A. A., & Revie, C. (2000). Thesauri on the web: Current developments and trends. *Online Information Review, 24*(4), 273–279.

Silva, I., Ribeiro-Neto, B., Calado, P., Moura, E., & Ziviani, N. (2000). Link-based and content-based evidential information in a belief network model. In *Proceedings of the 23rd annual international ACM SIGIR conference on research and development in information retrieval—SIGIR2000* (pp. 96–103). ACM Press.

Silveira, M. L., Ribeiro-Neto, B., Vale, R. F., & Assumpção, R. T. (2003). Vertical searching in juridical digital libraries. In *Proceedings of the 25th European conference on information retrieval research ECIR 2003* (pp. 491–501).

Smith, J. C. (1997). The use of lexicons in information retrieval in legal databases. In *Proceedings of the sixth international conference on artificial intelligence and law* (pp. 29–38). ACM Press.

STF (2002). Supreme Federal Court of the Federative Republic of Brazil. Available: www.stf.gov.br, 2002.

STJ (2002). Superior Court of Justice of the Federative Republic of Brazil. Available: www.stj.gov.br, 2002.

Tong, R. M., Reid, C. A., Crowe, G. J., & Douglas, P. R. (1987). Conceptual legal document retrieval using the RUBRIC system. In *Proceedings of the first international conference on artificial intelligence and law* (pp. 28–34). ACM Press.

ELSEVIER

Decision Support Systems 35 (2003) 113–127

Decision Support Systems

www.elsevier.com/locate/dsw

Enriching Web taxonomies through subject categorization of query terms from search engine logs

Shui-Lung Chuang*, Lee-Feng Chien

Institute of Information Science, Academia Sinica, Taipei 115, Taiwan

Abstract

In this paper, we propose a query-categorization approach to facilitating the engineering process of constructing Web taxonomies. One primary step in taxonomy construction is to acquire the domain-specific terminology terms and the mapping between the subjects and these terms. We introduce a technique for categorizing Web query terms from the logs of on-line search services into a predefined subject taxonomy based on their supposed popular search interests. The obtained experimental results show our technique's effectiveness in reducing the workload of human indexers in constructing Web taxonomies and also show its usefulness in various Web information retrieval applications.
© 2002 Elsevier Science B.V. All rights reserved.

Keywords: Information retrieval; Query categorization; Log analysis

1. Introduction

It is, undoubtedly, very valuable to put a huge amount of data into a well-organized taxonomy to help users perform many activities, such as searching for specific information and performing decision-making tasks. Hence, in recent years, many efforts have focused on facilitating the engineering process of taxonomy construction, e.g., the acquisition of subjects, the establishment of a subject hierarchy, the discovering of conceptual relationships, etc. [5,11, 13,14]. One primary step in this construction process is to acquire the domain-specific terminology terms and the mapping between the subjects and these terms.

The most common taxonomies available on the Web are the large directories used to organize Web sites (such as the one at Yahoo!). In order to enable such well-categorized resources to be easily processed by human indexers and retrieved by software programs (such as Internet search services), many domain-specific terms for each category and their relationships, e.g., synonym and hypernym/hyponym relationships, are required. However, most of the significant terms for Web information retrieval (IR) applications are proper nouns and new terms [19], both of which are hard to find in traditional hand-build knowledge bases, such as WordNet [10]. Manually collecting and organizing such term vocabularies for Web applications is neither practical nor cost-effective due to problems related to information updating, time consumption, and scalability. Thus, some automatic or computer-aided mechanisms are needed to make this process more efficient and adaptable.

* Corresponding author. Tel.: +886-2-27883799x1515; fax: +886-2-27824814.

E-mail addresses: slchuang@iis.sinica.edu.tw (S.-L. Chuang), lfchien@iis.sinica.edu.tw (L.-F. Chien).

0167-9236/02/$ - see front matter © 2002 Elsevier Science B.V. All rights reserved.
PII: S0167-9236(02)00099-4

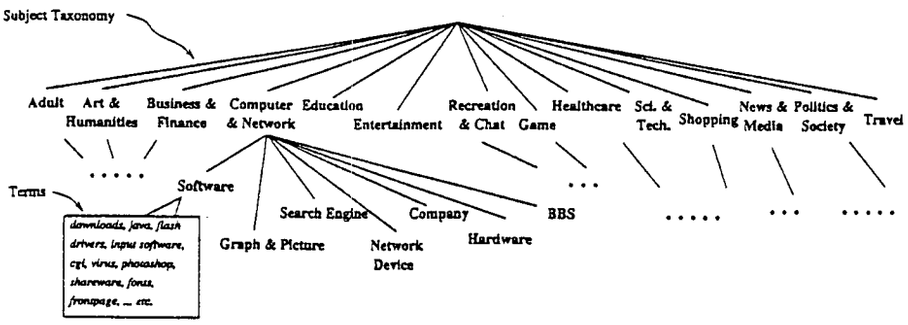


Fig. 1. An example of subject taxonomy with query terms.

Some of the excellent sources for obtaining Web significant terms are the query logs of on-line search services. In this work, we are interested in organizing such query terms[1] into a broader subject taxonomy (e.g., Fig. 1), where deeper analysis of domain-specific terminology and further discovery of term relationships can be performed under each subject domain. This can serve as a step toward enriching Web taxonomies. Thus, in this paper, our problem is as follows: If there already exists a predefined subject taxonomy, how can we place any given query terms into proper categories? For example, the appropriate category for the query term "Real Player" may be Computer/Company or Computer/Software Download.

To our knowledge, few reports in the literature have focused on the Web query term categorization problem. Recently, some document-based approaches to acquiring domain-specific semantic lexicons or to enriching the existing ontologies have been proposed [1,6,17]. Our work differs from these works in that we adopt the query logs as the terminology source, thus avoiding term segmentation and extraction problems commonly faced in document-based approaches. Some researches on term clustering are to a certain degree related to our research, such as the works on latent semantics, SVD, term relationship analysis, etc. [9,15,18]. Most of them have dealt with the automatic clustering of controlled indexed terms into clusters and with using them to assist IR systems in the query expansion process so as to improve the precision and recall ratios. There have been some works on terminology analysis for Web IR [2,4] and a similar work on the clustering of Web queries [3], in which a collection of user transactions from an Internet search engine are mined to discover clusters of similar queries. The clustered queries could then be used for term suggestion to help users form their own search requests. Our research differs from these works in that we try to structure each unknown query term into appropriate and meaningful subject categories, and provide possible corresponding subject domains for each unknown term.

To instantiate our research, the logs from three different search engines in different time periods were collected. Human efforts were applied to analyze these logs, and a two-level subject taxonomy with 14 major categories and 100 subcategories was constructed to express the popular search subject areas. Also, about 20 K of high-frequency query terms was categorized properly by human experts. This working environment then served as the basis for exploring use of the automatic query-term categorization approach.

Since a query term is too short to contain enough information in itself for automatic categorization, retrieved documents from real-world search engines are used in our approach to supplement the features of the candidate query term. Then, a classification method based on a term feature vector and a modified inner-product similarity measure is applied to this query term categorization task. Several experiments have been conducted to evaluate the performance of this approach. The top 1 inclusion rate, i.e., the rate of obtaining the top one category containing the most appropriate categories assigned by human indexers,

[1] Here the query terms we mention are the complete query strings submitted by users to search services; thus, those query strings that can be extracted from the logs. In the rest of this paper, the terms "query" and "term" refer to query terms when they are used without any additional explanation.

could reach 49.83%, and the top 5 inclusion rate could reach 78.64%. The obtained experimental results show that this approach is efficient in dealing with large numbers of queries and is adaptable to the dynamic Web environment. We believe that this approach will be also useful in the various Web IR applications aforementioned. An obvious application is that of collecting sufficient query terms for some sensitive resources, such as adults and drugs related materials, to improve the ability of a search engine to filter out sensitive queries.

In the rest of this paper, we will first introduce the classification method used to categorize query terms into a subject taxonomy. Next, the working environment in which we carried out the research will be introduced. This includes the query logs and the taxonomy we used in our experiments. Then, the conducted experiments and their results will be presented. Finally, we will discuss some possible applications and give a conclusion.

2. The query categorization problem

Because of the deep-hierarchical structures of real-world taxonomies, it is natural to hierarchically classify query terms into such taxonomies. That is, we first classify queries into top-level categories, and then for each category, we classify queries into its subcategories. Thus, the categorization of queries into a taxonomy is decomposed into a set of simpler problems, one at each node in the hierarchy, such that we only need to classify queries into a small set of categories. In the rest of this section, we will only discuss our approach to categorizing query terms into a set of categories.

Query categorization is defined as the problem of automatically assigning predefined categories to query terms according to the terms' supposed search interests or information needs. More formally, let C be a predefined taxonomy, and let $V=\{w_1, w_2, \ldots, w_n\}$ be the set of all query terms that have been categorized into C properly.[2] Query categorization is used to determine a category set $C(t) \subseteq C$ for a given unknown query term $t \notin V$, so each $c \in C(t)$ is considered as a possible category t is related to.

2.1. Classification model

A canonical way to an automatic classification[3] problem is to represent each candidate as a feature vector in which each dimension of the vector represents a candidate's attribute and its value indicates the candidate's quantity/quality for that attribute. Then, a classifier is constructed as a function that maps such an input feature vector to a numerical value indicating the confidence that the input candidate belongs to a class. The final class(es) can then be determined by means of some criteria, such as a predefined threshold value or some experimentally learned parameters. Thus, a classification problem can be solved by determining the proper features to be used and defining a classifier function to rank target classes.

Since a query term, usually containing one to two words in English [19] or two to three characters in Chinese [7], is too short to convey enough information in itself, some extra useful information is required to assist in the determination of corresponding categories. In our considered problem, a query term is a string submitted to a search engine to express a certain search request(s). Although the user's original search interest for a query term is hard to judge, the retrieved documents are believed to contain helpful information about the unknown query term. Collecting the required documents for each candidate query term can be easily performed by submitting the query string to on-line search engines. Thus, we assume that there exists virtually a document collection D, i.e., documents indexed by search engines, and we let D_t denote the set of documents that can be retrieved by query term t.

2.2. Extracting features from retrieved documents

It is helpful to understand the process a human indexer uses to determine the corresponding subjects

[2] A taxonomy is often organized hierarchically. In our work, we assume that any term assigned to an interior node really belongs to a leaf node that is a child of that interior node. Note that a term can still belong to multiple leaf nodes, but that all terms are only in leaf nodes and not interior nodes.

[3] Although the terms "classification" and "categorization" are definitely different in meaning, we will not differentiate between them in this paper and will use them interchangeably. The related terms "class" and "category," and "classify" and "categorize" will also be used in the same way.

of a given query term that is beyond his/her knowledge. From our observations, a human indexer may refer to the subject terms that occur in the documents retrieved by the candidate term when the subjects of the retrieved documents are hard to judge. Determining the subject(s) of the unknown term is then done based on the subjects of the subject terms in the retrieved documents. The proposed approach is, therefore, designed to simulate such human behavior.

The basic idea of our approach is to represent the feature vector of the candidate query term t by using the term features extracted from t's retrieved document set D_t. The feature space is defined by the precategorized query term vocabulary set V, in which each term has already been associated with proper category information. This approach is very similar to classifying a document based on the composed key terms in a conventional document classification process, or to tagging the part of speech for a word with its co-occurring neighboring words in linguistic analysis. Below, we will propose a ranking scheme used to estimate the confidence of a query term t belonging to a category c.

2.3. Tf–df-based category ranking

The approach employed in our work is to use term frequency (tf) and document frequency (df) information to give a weighted value for each occurring feature term. The intuition behind using the tf factor is that the degree that query term t belongs to category c is determined by whether there are many feature terms appearing in D_t with category c and by the number of times these feature terms appear. Also, feature terms appearing in more retrieved documents are assumed to be more relevant to t, which motivates use of the df factor. The formula is defined in the following.

Let n_w be the number of documents in D_t in which the feature term $w \in V$ appears, and let f_w be the raw frequency of term w in D_t (i.e., the sum of the number of times the term w occurs in the text of each document $d_i \in D_t$). The ranking function based on tf–df information is given by

$$R(t,c) = \sum_{w \in W_{t,c}} \frac{f_w}{\max_{k \in W_t} f_k} \log(n_w + 1), \qquad (1)$$

where W_t is the complete set of feature terms in D_t and $W_{t,c}$ is the set of those feature terms related to category c. Computing the maximum frequency over all feature terms is only done to beautify and make more reasonable the formula, and it does not have any effect if we only use the formula to rank a set of categories based on the same document set D_t. A count of one is added to n_w to avoid a logarithm value of zero.

3. The overall approach and working environment

In this section, we will describe the environment in which the research was carried out. The diagram depicted in Fig. 2 shows the overall concept of the

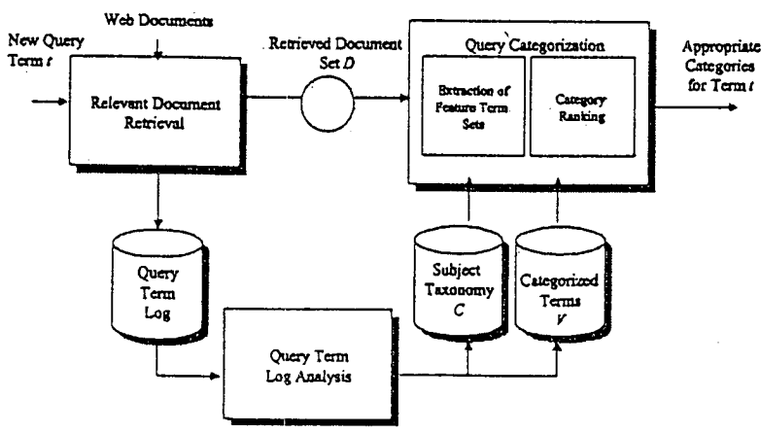


Fig. 2. An abstract diagram showing the concept behind the proposed approach.

SubjectCategorize(t, C, V, D, n)
t: the unknown query term
C: the set of subject categories
V: the set of feature terms
D_t: the highly ranked Web document set retrieved by term t
n: the number of target categories

1: **for all** $c \in C$ **do**
2: $W_{t,c} \leftarrow \emptyset$
3: **for all** $w \in V \wedge w$ appears in D_t **do**
4: calculate f_w (term frequency) and n_w (document frequency)
5: $W_{t,c} \leftarrow W_{t,c} \cup \{w\}$ where c is the category w belongs to
6: **for all** $c \in C$ **do**
7: $R(t,c) \leftarrow \sum_{w \in W_{t,c}} \frac{f_w}{\max_{k \in W_t} f_k} \log(n_w + 1)$
8: $C' \leftarrow \emptyset$
9: **for all** c in top n categories with highest $R(t,c)$ **do**
10: **if** c is a leaf node **then**
11: $C' \leftarrow C' \cup \{c\}$
12: **else** {c has children}
13: $C' \leftarrow C' \cup$ **SubjectCategorize**$(t, \{\text{children of } c\}, V, D_t, n)$
14: return top-ranked n categories in C' according to the decreasing order of $R(t,c)$

Fig. 3. An algorithmic procedure describing the hierarchical categorization process.

proposed approach, which is composed of three computational processes: query term log analysis, relevant document retrieval, and subject categorization. The function of query term log analysis, which will be further described in the following paragraphs, is to obtain the subject taxonomy and the categorized term set through analysis of search-engine logs. The relevant document retrieval process is performed to retrieve the most relevant document sets by working together with the search process of real-world engines. As we described above, this step is used to gather the features for each candidate query term. Finally, the categorization method described in the previous section is applied to determine the appropriate subject categories for each query term. Fig. 3 shows the detailed algorithmic procedure of the whole hierarchical categorization process, which recursively applies the categorization algorithm to the n, specified by the user, most possible candidate categories at each layer and, finally, returns the top-ranked n categories.

In the following paragraphs, we will describe in detail query term log analysis. This includes the logs we obtained from real search engines and the taxonomy we constructed for our experiment. Also, high-frequency query terms are manually analyzed and categorized into proper categories. We are interested in categorizing high-frequency query terms into popular search subject categories because these query terms are considered relatively stable foremost data for analyzing users' search interests or information needs. The obtained categorized query terms can then be treated as the basis for experimental evaluation.

3.1. Analysis of three log data sets

To instantiate the research, three query logs from the Dreamer[4] (D-1998), GAIS[5] (G-1999), and Openfind[6] (O-2000) search engines in Taiwan were collected as the basis for our analysis. The query logs contain a series of request entries, and each entry contains a query term, the IP address of the machine that sent the request, the corresponding timestamp, etc. Since we were only interested in the query terms themselves in our research, only the query-term parts of the logs were extracted.

Each query term in the logs was preprocessed by mapping the characters in its English part to lowercase, converting a sequence of spaces to only one, and

[4] http://www.dreamer.com.tw/.
[5] http://www.gais.cs.ccu.edu.tw/.
[6] http://www.openfind.com.tw/.

removing the spaces between Chinese characters. Furthermore, we removed those queries with only one byte or one Chinese character. No further processing was performed. To avoid significant preprocessing of query logs is to underscore the nature of the queries submitted by users.

Table 1 shows these three data sets based on the collection time periods, the counts of distinct query terms, the total query frequencies, and some basic statistics. In our research design, we focused on analyzing D-1998 and let G-1999 and O-2000 be cross-reference sets.

3.2. Structuring a subject taxonomy

To advance our study of the considered problem, we structured a popular subject taxonomy, a hierarchy of popular subject categories, to describe the subjects of users' search interests. As mentioned before, we are interested in the subjects the query terms correspond to. Instead of adopting the hierarchical structure commonly found in Web directories, such as Yahoo!, or in the classification schemes in library communities, such as the Dewey Decimal system, we used a bottom-up methodology by quickly reviewing the top 5000 query terms from D-1998 and then constructing the hierarchy based mainly on the analysis of human observations. The intention and information requests of each query term were estimated by several people with substantial experience surfing the Internet. Finally, a two-level hierarchical structure, consisting of 14 major categories together with 100 subcategories, was developed. Fig. 1 depicts a fragment of the whole taxonomy. The reason for structuring the taxonomy using such an approach was that we wanted it to reflect precisely the search interests in terms of locality of space and culture under which the queries were made.

Table 1
Statistics of three log data set

Data set name	D-1998	G-1999	O-2000[a]
Year	1998	1999	2000
Period	3 months	2 weeks	12 months
Number of distinct queries	228,566	114,182	3011
Number of total queries	2,184,256	475,564	2,493,211

[a] O-2000 collected the top 1000 query terms of each month in 2000, so it only contains 3011 distinct queries.

3.3. Manually categorizing high-frequency terms

Similar to the approach used to structure the taxonomy, categorization of high-frequency terms was performed manually by five Library and Information Science students together with a professional reference librarian for 3 months. These people had substantial experience surfing the Internet. In the whole process, each query term was examined, and the corresponding categories were determined according to subjective estimation of the information requests of the users who issued the query. For example, the query term "Microsoft" was categorized as an instance of the Company subcategory included in the major Computer category. Since it is very likely that a query term can have multiple information requests from different users, besides one major category, a secondary category was also assigned to each term if necessary for the purpose of cross-referencing. In this manual categorization process, a total of 18,017 terms with the highest frequencies from D-1998 were categorized properly. Though this set only represented 8% of the distinct queries, they totally formed 81% of the search requests in the test logs.

Admittedly, human classification is subjective even if there is interceder agreement between human indexers. However, in our study, the human analysts were involved mainly in classifying seed terms into appropriate subject categories. If there was an adequate number of feature terms in each category, we believe that the distortion caused by a few instances of invalid human categorization in fact would not seriously affect the accuracy of the auto-categorization approach.

3.4. Collecting retrieved documents

To collect the retrieved document set D_t for each query term t, we adopted Google[7] Chinese as the back-end engine. Each query term was submitted to Google, and then up to 100 search result entries were collected. The title and description of each entry were extracted as the representation of the corresponding document. In the final result, only 107 queries among 18,017 total queries had no retrieved documents.

[7] http://www.google.com/.

Table 2
Coverage comparison between D-1998 and G-1999

G-1999/D-1998	Top 1000	Top 20,000
Top 1000	583	914
Top 20,000	977	9709
All	992	14,721

3.5. Core term extraction

Many search requests are affected by ephemeral trends in querying, such as searches related to a newly released movie released or some specific events happening during the year. This situation really interests us. To determine the effects of time locality, query terms from D-1998 were matched and filtered using the G-1999 log to make a comparison of coverage. Table 2 shows the coverage comparison of query terms between the two data sets: D-1998 and G-1999. This table shows that 14,721 terms, nearly 77%, (the lower-right cell in the table) of D-1998's top 20,000 query terms still existed in G-1999's 2-week randomly selected log, which indicates that many core or important information requests were not much affected by time and were worthy of further study.

To obtain seed feature terms, query terms without the effects of time, so-called "core terms," are extracted based on whether or not the terms occur in both of the query logs. Except for some proper nouns like names of famous Web sites and people, an interesting finding was that core terms like "movie," "baseball," or "flight ticket" were found to mostly be subject terms. These core terms are considered to be more comprehensive and are often used by Web users to represent popular search interests and also by Web authors to indicate the key subjects of Web documents. Using core terms as features in the categorization process is believed to be more effective than just using common words.

4. Experiment

To assess the performance of the proposed approach, two tests using the data from D-1998 and O-2000 were conducted, respectively. The former was used to test the accuracy of the categorization approach compared with that of human analysis, and the latter was used to test the sustainability of different term sets like core terms or high-frequency terms as the feature set used for categorization.

The first test was performed to evaluate the performance achieved in categorizing the test query terms into our predefined two-level taxonomy with 14 major categories and 100 subcategories. The 9709 core terms were taken as the feature term set V. To ensure stability of the achieved performance, 20 test sets of query terms were prepared, each of which contained 1000 randomly selected noncore terms. The experiment was conducted using these 20 test term sets, and the performance achieved was the average of the results of these 20 trials. In addition, in order to reveal the effects of size variation of the core term set used as the feature set for categorization, we ran the experiment with different core term sets: the top 100, 200, ..., 900, 1000, 2000, ..., 9000, and 9709 terms from the seed feature term set, in accordance with the frequency of the core terms.

Some of the obtained correct rates are shown in Table 3, and the overall accuracy curve is depicted in Fig. 4, where top n means the highly ranked n candidate categories that contained the appropriate category. Also, Table 4 shows a set of randomly selected English query terms from our test sets (note that our query terms contain both Chinese and English queries) along with their human-assigned categories and machine-suggested top five categories. Note that

Table 3
Top 1–5 inclusion rates with various core term sets used as the feature sets

	100	300	500	1000	3000	5000	7000	9000	9709
(A) Inclusion rates for the top-level 14 categories									
1	36.80	46.70	48.20	51.90	55.40	59.50	60.00	59.90	60.20
2	48.10	62.20	61.80	67.40	72.60	76.60	77.30	78.00	77.90
3	53.50	68.70	70.50	74.60	82.40	84.40	84.60	85.40	85.70
4	56.60	73.60	75.90	81.60	87.10	88.90	89.20	89.60	89.50
5	57.90	79.00	80.20	86.80	90.80	90.90	91.30	91.60	91.70
(B) Inclusion rates for the second-level 100 categories									
1	24.98	36.30	38.79	43.67	46.77	48.77	49.40	49.57	49.83
2	31.00	46.12	49.49	55.40	61.51	63.87	64.58	65.26	65.55
3	33.28	50.23	54.29	60.91	67.82	70.28	71.36	72.09	72.34
4	34.83	53.39	57.44	64.73	71.62	73.78	74.70	75.78	75.87
5	36.06	55.58	59.76	67.34	74.50	76.67	77.42	78.47	78.64

The horizontal rows indicate the sizes of the core term sets used for categorization, and the vertical columns list the obtained inclusion rates for the top 1–5 candidate categories.

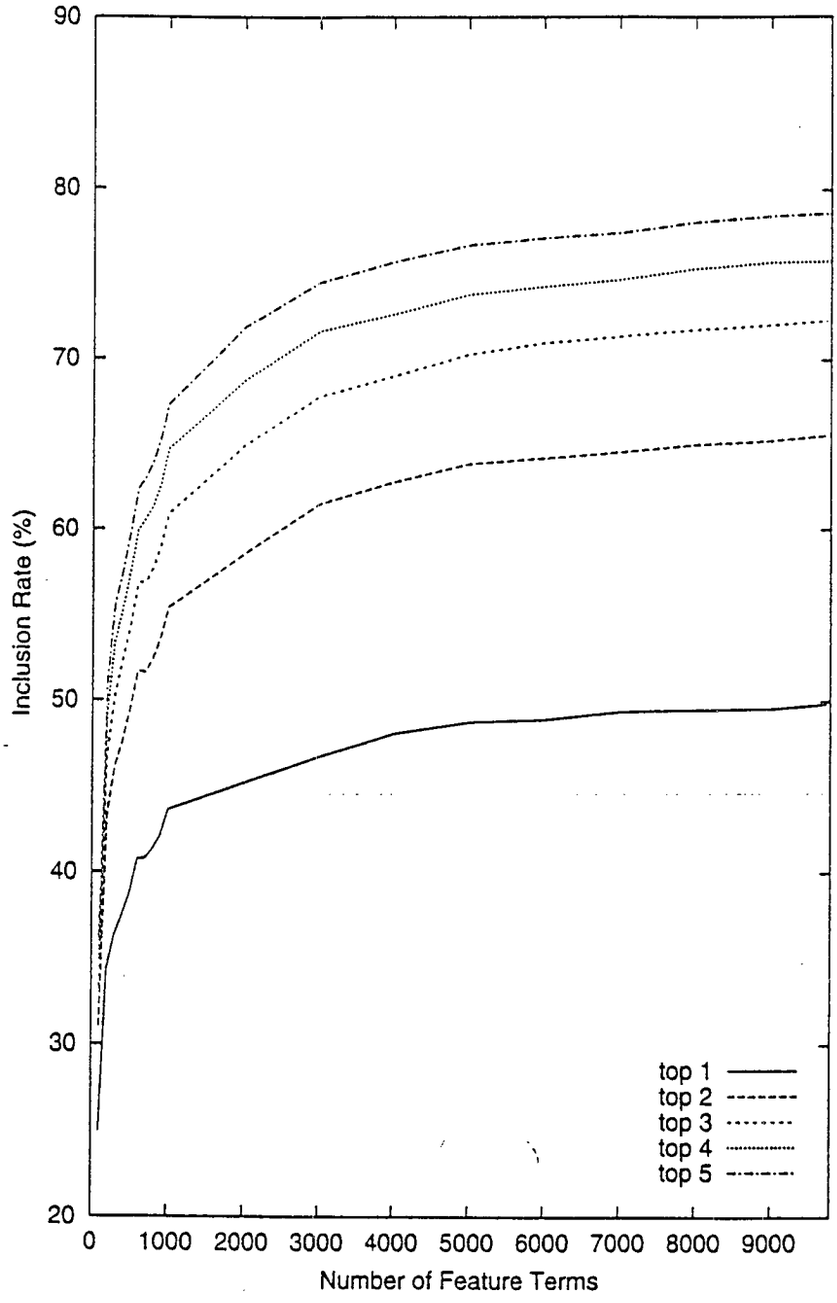


Fig. 4. Top 1–5 inclusion rates with the core term set for second-level categories.

the top five categories are shown in the order of their R values, i.e., the one on the left-hand side has a larger R value than the one on the right-hand side does.

Even with the simple tf–df ranking scheme, the performance seems to be acceptable. Considering only the top 1 categorization result for second-level categories, the average top 1 inclusion rate, i.e., the rate of the obtained top one category containing the most appropriate categories assigned by human indexers, is 49.83%. If we consider the performance of the top five categories, the inclusion rate can reach 78.64%. The achieved performance is assumed nearly compet-

Table 4
An example of query terms with their top five suggested categories

Query term	Corr. cat.	Suggested five cat.
real player	cd,cn	*cd* *cn* mn ch pc
star trek	en	ei gg *en* cd ds
michael jackson	ei	em cd tl en cn
tennis	fs	*fs* e3 cd ds gg
case tool	cd	ch ds *cd* e3 ks
erotica	ss	cn *ss* mn cd cs
meg ryan	ei	en cn cd em cp
i phone	cd,cn	*cd* lp tl *cn* e3
quarkxpress	cd	*cd* kb cg aa mn
doom2	gg	cn cd *gg* ds cb
australia	tf	tl *tf* cd cn ds
fishing	fs	tl cn e3 *fs* cd
chinatrust	bb	bm tl tf *bb* cn
geocity	cn	*cn* cd mn cs ds
motorora	lp,bn	cn *lp* *bn* dt cd
age of empires	gg	*gg* cd cn ch cg
photo shop	cd	*cd* cp dt cn kb
airplane	tp	*tp* bf cn cd mn
cafe	le	*le* cn mn cd em
nokia8810	lp	cn cd *lp* ds ls

Major category	/ Sub-category	: ID
Art&Humanities	/ Art	: aa
Business&Finance	/ Bank	: bb
	/ Information	: bf
	/ Money	: bm
	/ Telcom	: bn
Computer&Network	/ Download	: cd
	/ Group	: cg
	/ Hardware	: ch
	/ Network	: cn
	/ Picture	: cp
	/ Search Engine	: cs
Education	/ Group	: dg
	/ School	: ds
Entertainment	/ MP3	: e3
	/ Individual	: ei
	/ Music	: em
	/ Movie	: en
Recreation&Chat	/ Entertainment	: fe
	/ Sport	: fs
Game	/ Game	: gg
Sci.&Tech.	/ Book	: kb
	/ Science	: ks
Shopping	/ Phone	: lp
	/ Shopping	: ls
News&Media	/ News	: mn
Adult	/ Sex	: ss
Travel	/ Local	: tl
Travel	/ Plane	: tp

itive with that of human indexers. Human indexers normally assign one or two categories for each test query term due to the difficulties involved in complete analysis. It is worth noting that although many of the top five categories are not exactly the same as the categories assigned by human indexers, some of them are still related to the human-assigned categories and cannot simply be considered as miscategorized. For example, the query "michael jackson" shown in Table 4 was assigned to Entertainment/Individual by humans. Although this human-assigned category does not appear in the machine-suggested top five categories, the suggested category Entertainment/Music is surely related, and the category Computer/Download may also be considered related because someone may want to download the singer's music files. Therefore, the suggested top categories are found to be rather helpful in many respects; for example, they can be used by human indexers to reexamine the correctness of their assigned categories.

Our second test was conducted to compare the sustainability of various feature term sets for categorization. We used O-2000 as our test set, which had a 2-year lag compared to the feature term set from D-1998 in the first test. Among the total of 3011 distinct query terms, 1265 terms were not found in the term set that had been categorized. These unlabeled query terms were then categorized by the human indexers as described in the previous section, and the result was treated as the testing set. To compare the effects of various feature sets, we adopted four feature sets selected based on D-1998: core terms (coreterm), high-frequency terms (freqterm), terms randomly selected from among high-frequency terms (freqrand), and terms randomly selected from among 18,017 categorized terms without considering their frequen-

cies (baseline). The experiment was conducted in the same way as the first one was.

The results are shown in Fig. 5. The baseline set obviously had the poorest performance. The coreterm set and freqterm set achieved comparable results. While the freqterm set outperformed the freqrand set, the figure shows that the total frequencies (ranks) of the feature terms greatly influenced the performance. Although the experimental results do not strongly support our previous assumption that core terms perform better than frequency terms in the categorization process, it can be assumed that better performance would be achieved if longer time lags were considered.

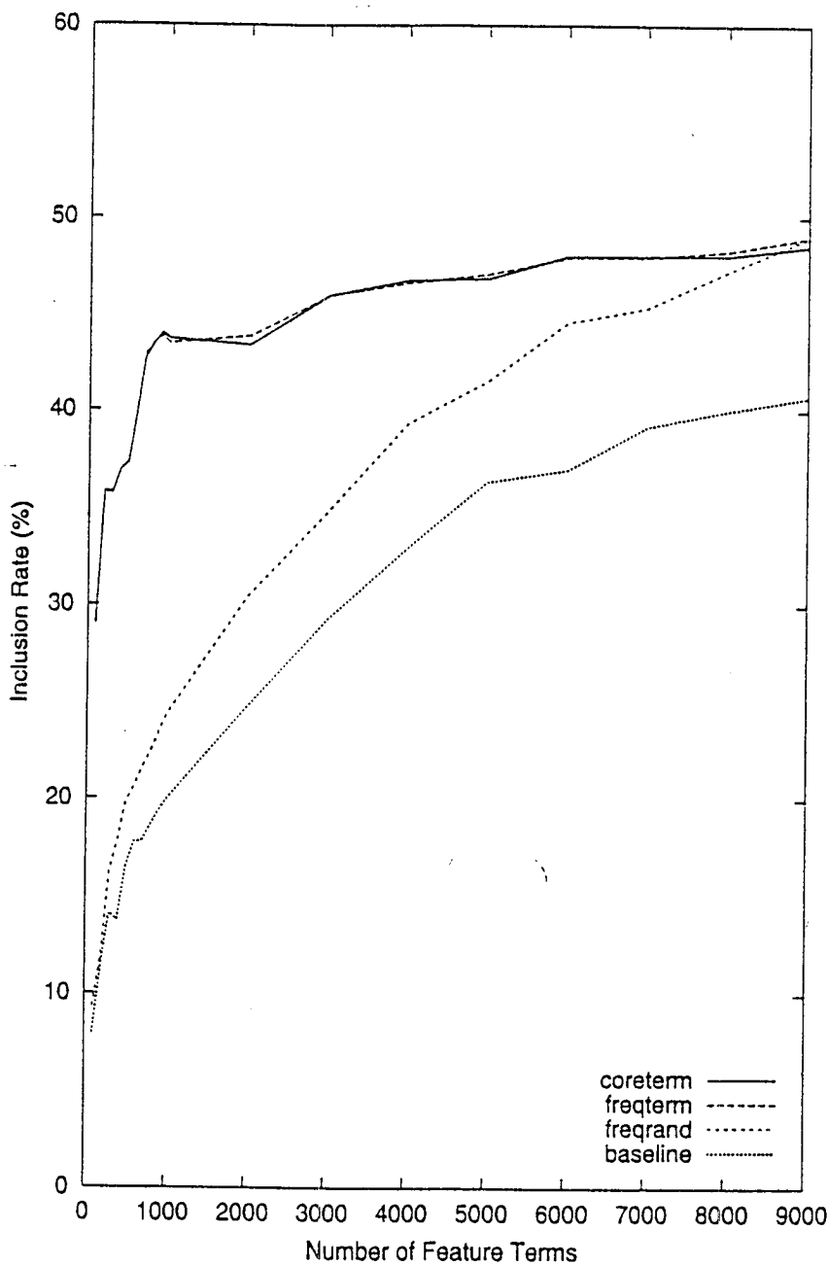


Fig. 5. Top 1 inclusion rates for four different feature term sets.

4.1. Discussion

One of the factors that causes the robustness of the proposed approach to increase is the representative and coverage of the core terms as the basic feature terms in this categorization task. These core terms were extracted based on adequate logs and human effort. It is worth noting that the top 1 inclusion rate could reach 24.98% only using the top 100 core terms as the feature set, compared with a rate of 7.92% obtained using 100 randomly selected feature core terms (cf. Fig. 5). This reveals that these top core terms might represent some major search interests, and that they are effective features for categorizing query terms into certain categories. In fact, Fig. 4 shows that a small set of precategorized core terms, e.g., 1000–3000, was useful for achieving an acceptable performance. This number is much smaller than the total number of query terms. However, there were weaknesses in our initial study.

One of the factors that may weaken our approach is the weak representation of corresponding retrieved documents for test query terms. We represent each document using only the title and description extracted from the top 100 search results from search engines. The titles are normally useful. However, the descriptions are usually autogenerated by the search engines and, thus, are often not precise or meaningful. One possible alternative way is to use the document contents or the anchor texts of that page, which are texts associated with links that point to that page. If this kind of representation is used, it will be necessary to run a Web crawler to download the corresponding pages. This requires lots of band width and storage space, so this approach is not suitable under most environments. In addition, due to the hypertext design principle, the web page content may also be not representative enough. Anchor texts are sometimes precise descriptions for a page, but they are usually short and must be interpreted in their context. Thus, using anchor texts as the representation of a document may make it too narrow for obtaining a sufficient feature term set.

One other reason is the ambiguous nature of terms. As we have mentioned, query terms are usually very short. This means that some terms have multiple information requests. Comparing the performance of top 3 or 5 with that of top 1, there is obvious improvement. This shows that our approach to multicategorization is still effective. However, determining how many categories should be chosen has not yet been explored.

Analyzing the instances of erroneous categorization of query terms by the machine for each subject category, it is noted that some categories tend to have more ambiguous or annoying terms than others, such as the Computer and Network category. These terms usually have little discrimination value for categorization and prevent correct category assignment for each term. For example, many Web documents contain some computer-related terms like "search," "Web," and "homepage;" hence, it is very likely that query terms containing these feature terms in their retrieved documents will be assigned to the Computer and Network category. In other words, these annoying terms act like stop words and need to be filtered out from the feature term set to prevent inappropriate categorization.

Furthermore, each category poses different problems that need to be solved. For example, personal name identification techniques are useful for dealing with the Entertainment/Individual subcategory since query terms belonging to this category are mostly personal names. Considering the sample query terms "michael jackson" and "meg ryan" listed in Table 4, they were not correctly categorized into the Entertainment/Individual category. However, they are really suggested into related Entertainment subcategories.

In addition, some other factors may affect the correctness of category assignment and need to be further studied, such as the relevance and numbers of the retrieved documents, and the sufficiency of the surrogates of these documents as described in the previous paragraphs. Also, as a task of categorization into hierarchical taxonomies, there are still many approaches that need to be explored to better utilize the hierarchical structure, such as that of applying deeper analysis of terms to terms and of terms to subjects to determine flexible features for each interior node in the taxonomy [12].

The experimental results are promising that various new terms given by users can be categorized automatically or semiautomatically if high categorization accuracy is required. To make our approach adaptive, we only need to apply human effort periodically to examine whether the initial feature terms are appropriate or not and perform updating if necessary. Using

this human–machine integrated approach, terms can be categorized properly.

5. Discussion of possible applications

Our term categorization approach for real-world query terms may be useful in some applications.

5.1. Live thesaurus construction

Query term categorization serves as a first step in the construction of a thesaurus for Web search. For query terms categorized in the same classes, more accurate relationships between terms such as abbreviations, synonyms, related terms, broader terms, narrow terms, and translations might be further discovered. Such thesaurus information can be very useful in some applications, such as term suggestion for interactive Web search [8].

Our work provides a good beginning for constructing such a Web thesaurus. Through the term categorization process, terms can be grouped into several meaningful subject categories. With subject information of each term, some kind of term suggestion can be performed. Furthermore, if we think a category is still too rough to convey a strong relationship between terms, we can apply some clustering techniques to terms in the same category. This will reduce the size of the term set for the clustering algorithm and, of course, enable it to more efficiently and reliably find really relevant terms without disturbance from other irrelevant terms.

5.2. Sensitive query filtering

To prevent commercial search engines from providing inappropriate information to certain users like children, it is necessary to have a query term filter for filtering out sensitive query terms, such as pornography-related terms. The query term filter problem is similar to the search term categorization problem: the problem is to determine whether a given term can be categorized as porn-related or not. Our approach directly provides a possible solution to this problem. Table 5 shows some other statistics of the categorization with the first-level 14 major categories from our previous experiment described in Section 4. Recall that the experiment was conducted 20 times on 1000 randomly selected terms of each time for a total of 20,000 trials. Each table cell shows, for one category, the recall rate, the rate of terms in that category being correctly categorized, and the precision rate, the rate of terms categorized in that category really belonging to that category. The results show that the proposed method can be applied to query term filtering within such categories as Adult, Computer, etc. The proposed approach is useful for collecting users' query terms in these domains. Combined with final examination by humans, a collection of query terms can be incrementally updated. A real-world query term filter for filtering out pornography-related terms for Web image search has been successfully developed, and more than 20,000 sensitive query terms have been collected using the proposed approach.

5.3. Observation of Web users' search interests [16]

Information about the needs of Web users is valuable for enhancing the effectiveness of search engines and Web content organization. The query term logs of search engines are considered to be the most useful data for monitoring users' information needs and search interests. With our approach to subject categorization of query terms, it is a straightforward task to construct an automatic system for Web search engines

Table 5
Recall/precision statistics of categorization with the first-level 14 major categories

	Humanities	Business	Computer	Education	Entertainment	Chat	Game
Top 1	43.41/70.10	70.38/62.41	80.11/53.29	72.28/76.48	55.36/75.86	57.35/63.71	65.51/77.07
Top 3	71.26/40.39	91.81/28.15	96.07/22.35	92.16/43.06	80.44/46.61	85.60/30.81	87.69/51.66
	Health	Science	Shopping	Media	Society	Adult	Travel
Top 1	70.07/77.22	43.10/43.43	60.60/60.95	54.74/49.87	41.21/78.75	51.26/81.02	74.51/56.63
Top 3	86.98/54.41	74.20/12.07	91.45/21.58	83.03/20.86	78.17/38.61	75.92/65.49	90.40/21.53

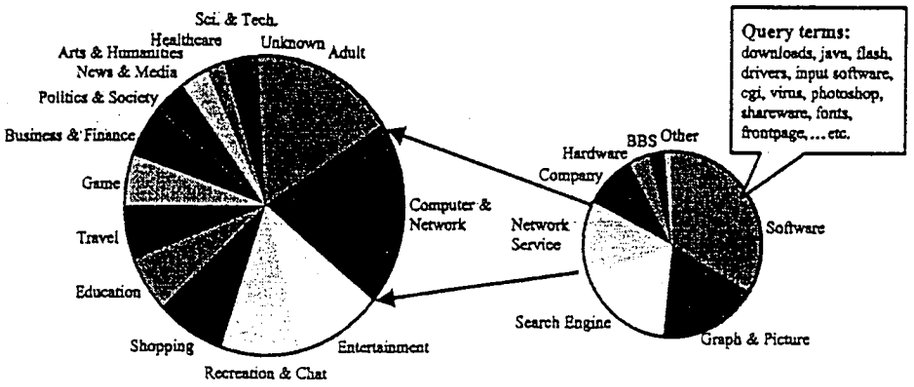


Fig. 6. An illustration of an information request distribution for some subject categories.

or digital library systems to monitor changes in users' search requests, including the distribution of users' search subject categories, and the frequencies of the query terms in each category. To show this idea more clearly, the distribution of information requests of some subject categories from the analysis of D-1998's log is depicted in Fig. 6. Among the 14 major categories, the Computer category has the highest percentage of queries, 19.9%, followed by Adult at 16.9%, Entertainment at 9.8%, and so on. Each major category is also decomposed by related subcategories. Obviously, an up-to-date distribution of the information needs corresponding to each subject category can be easily obtained using our approach.

5.4. Web page classification/filtering

The problem of classifying Web pages is undoubtedly very important in today's Web environment, and it also presents a big challenge to the traditional text classification approaches. When we consider the determination of the class of a document made by a human being, we can obviously observe that the determination of the target class is usually dependent on only a few dominate key terms. When conventional machine learning approaches are applied to some corpora, the effect of such key terms may been decreased by other unrelated terms. This phenomenon is more obvious in the Web page environment. Imagine the difficulty of classifying a company's front page containing only a few images or animations. In such a case, the traditional term-based approach can only depend on the page title or anchor texts from other pages pointing to the candidate page. Such contents are usually very brief, and only one or two key terms can be used to determine the class of the pages. Search logs provide us with a good source of vocabularies used in Internet communities. By associating each term with some subject information using our approach, we may be able to form these query terms into a good feature set for classifying Web pages. Because this is surely an another big problem and much work needs to be done, we only point out this idea here.

6. Conclusion remarks

In this paper, we have presented an approach to categorizing query terms that are obtained from on-line search engine logs into broader subject categories of taxonomies for use in Web applications. The approach successfully integrates human and machine efforts. First, a set of high-frequency core query terms along with their subject information can be determined by human experts. With this seed core term set, an automatic categorization process can be developed for categorizing each newly given query term. The experimental results show that various new terms given by users can be categorized automatically, and this gives that that our approach can be applied as a good beginning for constructing Web taxonomies or for enriching existing ones using query terms. Other applications that can benefit from our work have also been pointed out.

References

[1] E. Agirre, O. Ansa, E. Hovy, D. Martinez, Enriching Very Large Ontologies using the www, ECAI 2000, Workshop on Ontology Learning, Berlin, Germany.

[2] P.G. Anick, S. Tipirneni, The paraphrase search assistant: terminological feedback for interactive information seeking, Proceedings of the 22nd ACM International Conference on Research and Development in Information Retrieval (SIGIR'99), Berkeley, USA, August 15–19, 1999, ACM Press, New York, USA, 1999, pp. 153–159.

[3] D. Beeferman, A. Berger, Agglomerative clustering of a search engine query log, Proceedings of the Sixth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, Boston, MA, USA, August 20–23, 2000, ACM Press, New York, USA, 2000, pp. 407–416.

[4] P. Bruza, R. McArthur, S. Dennis, Interactive internet search: keyword, directory and query reformulation mechanisms compared, Proceedings of the 23th ACM International Conference on Research and Development in Information Retrieval (SIGIR'2000), Athens, Greece, July 24–28, 2000, ACM Press, New York, USA, 2000, pp. 280–287.

[5] R. Byrd, Y. Ravin, Identifying and extracting relations in text, Proceedings of the Fourth International Conference on Application of Natural Language to Information Systems (NLDB'99), Austria, June 17–19, 1999.

[6] L.-F. Chien, Pat-tree-based adaptive keyphrase extraction for intelligent Chinese information retrieval, Information Processing & Management 35 (1999) 501–521.

[7] L.-F. Chien, H.-T. Pu, Important issues on Chinese information retrieval, Computational Linguistics and Chinese Language Processing 1 (1) (1996) 205–221.

[8] S.-L. Chuang, H.-T. Pu, W.-H. Lu, L.-F. Chien, Auto-construction of a live thesaurus from search term logs for interactive web search, Proceedings of the 23rd Annual International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR2000), Athens, Greece, July 24–28, 2000, ACM Press, New York, USA, 2000, pp. 334–336.

[9] S. Deerwester, S.T. Dumais, G.W. Furnas, T.K. Landauer, R.A. Harshman, Indexing by latent semantic analysis, Journal of the American Society for Information Science 41 (6) (1990) 391–407.

[10] C. Fellbaum, Wordnet: An Electronic Lexical Database, MIT Press, Cambridge, 1998.

[11] U. Hahn, K. Schnattinger, Towards text knowledge engineering, Proceedings of the 15th National Conference on Artificial Intelligence (AAAI'98), Madison, Wisconsin, USA, July 26–30, 1998, AAAI Press, Menlo Park, California, USA, 1998, pp. 524–531.

[12] D. Koller, M. Sahami, Hierarchically classifying documents using very few words, Proceedings of the Fourteenth International Conference on Machine Learning (ML'97), Nashville, Tennessee, USA, July 6–12, 1997, Morgan Kaufmann, San Francisco, USA, 1997, pp. 170–178.

[13] A. Maedche, S. Staab, Discovering conceptual relations from text, Proceedings of the 14th European Conference on Artificial Intelligence (ECAI-2000), Berlin, Germany, August 20–25, 2000.

[14] A. Maedche, S. Staab, Mining ontologies from text, Proceedings of the 12th International Conference on Knowledge Engineering and Knowledge Management (EKAW-2000), Juanles-Pins, French Riviera, France, October 2–6, 2000, Springer-Verlag, Heidelberg, Germany, 2000.

[15] R. Mandala, T. Tokunaga, H. Tanaka, Combining multiple evidence from different types of thesaurus for query expansion, Proceedings of the 22nd ACM International Conference on Research and Development in Information Retrieval (SIGIR'99), Berkeley, USA, August 15–19, 1999, ACM Press, New York, USA, 1999, pp. 191–197.

[16] H.-T. Pu, S.-L. Chuang, Auto-categorization of search terms toward understanding web users' information needs, Proceedings of the 3rd International Conference of Asian Digital Library (ICADL2000), Seoul, Korea, December 6–8, 2000.

[17] E. Riloff, J. Shepherd, A corpus-based approach for building semantic lexicons, Proceedings of the Second Conference on Empirical Methods in Natural Language Processing, Providence, Rhode Island, USA, August 1–2, 1997.

[18] M. Sanderson, B. Croft, Deriving concept hierarchies from text, Proceedings of the 22nd ACM International Conference on Research and Development in Information Retrieval (SIGIR'99) (1999) pp. 206–213.

[19] C. Silverstein, M. Henzinger, H. Marais, M. Moricz, Analysis of a very large altavista query log, DEC SRC Technical Note, 1998.

Shui-Lung Chuang is a research assistant at the Institute of Information Science, Academia Sinica, Taiwan, for his military service since October 1999. He received his BS and MS degrees in Computer Science and Information Engineering from the National Taiwan University in 1997 and 1999, respectively. His master thesis was on information extraction from Web resources, and won the Best Thesis Award from ACM Taipei Chapter in 1999. Currently, his research interests include Web information retrieval and mining, knowledge acquisition and organization, text classification, and intelligent agent.

Lee-Feng Chien received his Computer Science PhD at the National Taiwan University in 1991. Since 1993, Dr. Chien joined in the Institute of Information Science, Academia Sinica, and is currently an associate research fellow of the Institute. Dr. Chien's research interests include information retrieval, Web mining, spoken language processing, and natural language processing. In these areas, he has published over 100 technical articles in scientific journals and conference proceedings such as IEEE Trans SAP, Computational Linguistics, JASIS, IP&M, ACM SIGIR, ICASSP, COLING, ACL, WWW, etc. Dr. Chien is currently an associate editor of ACM Transaction on Asian Language Information Processing. Besides, he usually served as a committee member in international conferences, such as a PC member of ACM-SIGIR'99, 00, 01, IRAL'97 00, ICCPOL'95 99, and the PC chair of the 1999 International Workshop on Information Retrieval with Asian Languages (IRAL'99). He was a co-recipient of the 1998 ACM SIGIR Best Poster Presentation Award for Late-breaking Research in Melbourne and a co-adviser of the recipient of the 1999 ISCSLP Best Student Paper Award in Singapore. He was also the owner of the 1997 K.T. Li Award by ACM Taipei Chapter for his contribution in Chinese Information Retrieval.

Int. J. Human-Computer Studies (2000) 53, 741–763
doi:10.1006/ijhc.2000.0416
Available online at http://www.idealibrary.com on IDEAL®

Evaluating the effectiveness of visual user interfaces for information retrieval

A. G. SUTCLIFFE
Centre for HCI Design, Department of Computation, UMIST, P.O. Box 88, Manchester M60 1QD, UK

M. ENNIS
Centre for HCI Design, School of Informatics, City University, Northampton Square, London EC1V 0HB, UK

J. HU
Centre for HCI Design, Department of Computation, UMIST, P.O. Box 88, Manchester M60 1QD, UK

(Received 31 January 2000, and accepted in revised form 31 May 2000)

An integrated visual thesaurus and results browser to support information retrieval was designed using a task model of information searching. The system provided a hierarchical thesaurus with a results cluster display representing similarity between retrieved documents and relevance ranking using a bullseye metaphor. Latent semantic indexing (LSI) was used as the retrieval engine and to calculate the similarity between documents. The design was tested with two information retrieval tasks. User behaviour, performance and attitude were recorded as well as usability problems. The system had few usability problems and users liked the visualizations, but recall performance was poor. The reasons for poor/good performance were investigated by examining user behaviour and search strategies. Better searchers used the visualizations more effectively and spent longer on the task, whereas poorer performances were attributable to poor motivation, difficulty in assessing article relevance and poor use of system visualizations. The bullseye browser display appeared to encourage limited evaluation of article relevance on titles, leading to poor performance. The bullseye display metaphor for article relevance was understood by users; however, they were confused by the concept of similarity searching expressed as visual clusters. The conclusions from the study are that while visual user interfaces for information searching might seem to be usable, they may not actually improve performance. Training and advisor facilities for effective search strategies need to be incorporated to enhance the effectiveness of visual user interfaces for information retrieval.
© 2000 Academic Press

1. Introduction

Few empirical studies of visual user interfaces (UIs) have focused on the strategies that users employ when searching through or interacting with complex images. We do know that the mapping between visual representations and the user mental model is important for promoting comprehension and effective use of visual UIs, but there is little

1071-5819/00/110741 + 23 $35.00/0
© 2000 Academic Press

information on how visual structures provide cues for learning. One solution is to provide users with facilities for organizing their own maps of information spaces. Usability studies have shown that self-organization of visual mental maps is effective (Czerwinski, Van Dantzich, Robertson & Hoffman, 1999); however, in many cases the user may have no starting point for self-organization, so the system needs to present a learnable image that portrays the structure of an underlying system or database. While many designs have been developed to provide visualizations of hierarchies, networks and time lines, evaluation studies are less common. Some metaphors have demonstrated effectiveness, such as the time line metaphor (Plaisant, Milash, Rose, Widoff & Shneiderman, 1996), and simple categorial grouping of results in information retrieval searches (Pirolli, Schank, Hearst & Diehl, 1996). While a development framework for design of visualizations to match the users' task and underlying system structures has been proposed (Card, Mackinlay & Shneiderman, 1999), questions still remain about the effectiveness of visualizations in specific formats or metaphors in supporting particular tasks.

In our previous work, we investigated user interaction with 3D information browsing displays (Sutcliffe & Patel, 1996) and demonstrated that 3D images do not appear to provide a significant advantage over two-dimensional displays. We also investigated the strategies users employed while interacting with 3D information visualizations and showed that different patterns of visual search range from systematic searches following the image structure to random sampling. In subsequent studies on standard GUI-based information-retrieval systems we found that users' search strategies were a major determinant of search performance and this seemed to over-ride usability problems (Sutcliffe, Ryan, Springett & Doubleday, 2000). One of the major determinants of search success was the user's persistence in iterative cycles of search and careful evaluation of retrieved articles to determine their relevance. The other major factor was choice of appropriate keywords. Furthermore, search facilities needed to be closely coupled so users could dynamically explore the relationship between query terms and the retrieved documents. The effectiveness of close coupling has been demonstrated in the alphaslider system (Ahlberg & Shneiderman, 1994), which integrated visualization and interactive querying; however, these designs did not support query articulation or exploration of meta-data, and were limited to query templates with value ranges. Other prototypes have provided integrated support for both query formulation and results evaluation, e.g. in the OKAPI system (Hancock-Beaulieu, Fieldhouse & Do, 1995) relevance feedback facilities allow the user to extract terms from retrieved documents for reuse in subsequent queries. Although a wide variety of visual user interface designs have been developed for browsing thesaurus and classification structures, visualization of retrieved results is less common. Furthermore, visualization for query formulation and results evaluation does not seem to have been integrated in one system. This forms the starting point for the design we report in this paper.

As well as building a novel visual interface for information searching we were interested in how visual metaphors represent system models to the user. Most visualizations have represented information structures as hierarchies or networks of various forms (Card, Robertson & Mackinlay, 1991). More adventurous representations such as data walls portray design concepts such as filters and context of focus, while tilebar browsers have depicted properties of retrieved documents. In this study, we attempted to

test how visualization might be able to communicate the underlying search functionality of a system. We chose similarity-based searching with the latent semantic indexing algorithm (Landauer & Dumais, 1997) and cluster analysis of retrieved results. This gives a more sophisticated search process than the traditional search engine with keywords and relevance ranking based on goodness-of-fit metrics. The question was how well could visualizations communicate something of the search engine process as well as the structure of the information base.

The paper is organized into five sections. In the next section, we describe the design of the integrated information browser system. This is followed by the methods used in the empirical study on the design to determine its effectiveness and analyse usability problems. We then report performance results and usability problems that were experienced with the design; and users' behaviour and search strategies. The paper concludes with a discussion of lessons learned from the empirical study.

2. Interface design

The Integrated Thesaurus-Results Browser was designed to support cycles of iterative querying, browsing and evaluation of retrieved results. In previous studies (Sutcliffe & Ennis, 1998), we proposed a cognitive task model that described user behaviour during different phases of information searching tasks and the search support facilities required to support each phase. For instance the model indicated that the early phases of articulating search needs and forming queries should be supported by thesauri and term suggestion facilities, while the evaluation phase needed visualization of retrieved results, with sorting and clustering to show grouping and relationships between documents, as well as relevance ranking of articles. Furthermore, our model indicated a close coupling between visualization in the query formulation and results evaluation. These requirements have been discovered and implemented before in isolation, for instance, the scatter/gather browser (Pirolli *et al.*, 1996) provides clustering algorithms for grouping results sets and limited visualization, while the information visualizer (Card *et al.*, 1991) concentrates on more sophisticated visualization for the query articulation phase.

The Integrated Thesaurus-Results Browser takes these design concepts one stage further. The screen layout is illustrated in Figure 1.

A tiled window layout was adopted to increase the ease of cross referencing between queries, meta-data representation in the thesaurus, visual summaries of the result sets and viewing documents. A tiled display saves the user work and reduces working memory load by allowing the user to continually scan all the necessary information for the task in hand. Inevitably, the downside of this design choice is a more busy display and limited area for viewing large-scale visualizations. The screen is divided into six areas. The top left-hand panel contains the query formulation dialogue, which allows query terms to be entered directly or to be selected from the thesaurus immediately below it. The thesaurus has a standard hierarchical structure, ranging from general to more specialized terms, although synonyms and related terms were not included in the case study system. The thesaurus hierarchy contained 118 terms arranged in six top level categories which expanded into 2–3 lower levels. Since the whole thesaurus was too large to fit into the allocated space, controls allowed the user to expand sub-trees by single clicking upper-level terms. Double-clicking terms entered the term selected as part of the

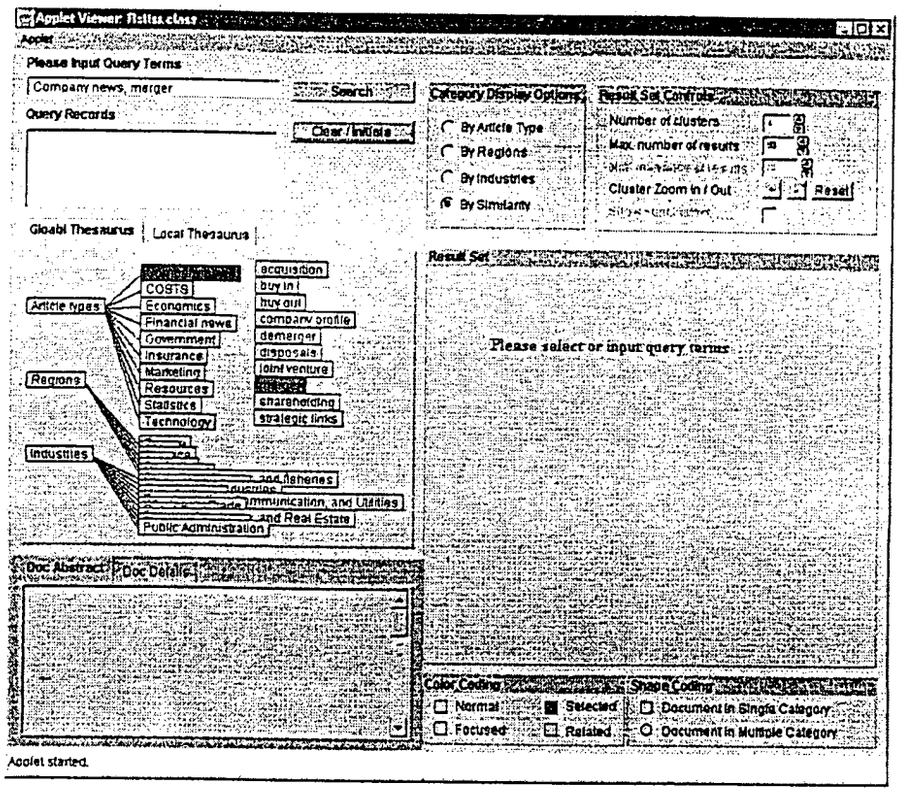


FIGURE 1. Overview of the Integrated Thesaurus-Results Browser system. The user selects terms from the thesaurus (middle left) to form the query (top left). The LSI search algorithm followed by cluster analysis produces the bullseye display (middle right) that gives groups of similar results in each cluster ranked by relevance to the user's query in the target "bullseye" rings. Filters and controls in panel (top right) enable the user to control the number of clusters or selected results grouped by indexing categories. Selecting a document from the bullseye displays the abstract and document details in the results viewer (lower left).

current query. The bottom part of the left side of the screen contained the abstract and details of the retrieved document, which were selected in the results browser on the middle right-hand side of the screen. The results browser contained one or more "bullseye" cluster displays for similar documents with the retrieved results for a query. The bullseye metaphor encoded two properties of the results: first, relevance was represented by rings with higher relevance in the centre of the bullseye, moving outwards to lower scores; secondly, similarity between documents was expressed by clusters on the browser.

The system used the latent semantic indexing (LSI) algorithm (Landauer & Dumais, 1997) as its search engine. LSI works by calculating the similarity between word distributions in two or more documents using an eigenvector algorithm. We used it first to retrieve documents which matched the query keywords using the LSI term to document similarity calculation as a relevance measure. Individual documents within a retrieved set were then matched for similarity using LSI's document to document similarity. In the latter, LSI calculates the co-occurrence between the distribution of all

terms in any two documents, giving a similarity score for each pair of documents. Similarity scores were calculated for all dyadic combinations of documents in a set. A minimum spanning tree cluster analysis algorithm was then applied to the matrix of document to document scores to discover groups of highly related articles. The cluster analysis works by selecting the most closely related document pair, working downwards from those documents to discover nearest neighbour closely related documents and so on. Branching in the tree depends on the number of child documents related to the current "parent". Clusters can be extracted at different levels of similarity score by counting all documents connected to a parent at an arbitrary level.

In the results browser the default cluster display was set at two groups, although this was under user control, and in the case of a tie at the first level (e.g. 3 equal scores for the first set of children documents), the system defaulted to a single group. The number of clusters selected could be set by the user with the control panel on the top right hand of the screen. This also allowed clusters to be selected by a standard indexing method, so documents were grouped by any category used to index the database. The control panel also provided parameters to set the number of results retrieved in relevance order, percentage relevance cut-off, etc.

The results in the bullseye cluster display appear in a spiral from the centre outwards with distance from the centre encoding relevance to the user's query. The bullseyes can be moved and expanded to improve the view when dense clusters of documents are displayed. Moving the cursor over each document symbol causes the short title to be displayed as "hover text" and double clicking on the icon triggers display of the abstract and document details in the results viewer area, as illustrated in Figure 2.

The system was implemented on a sub-set of the Financial Times McCarthy database of newspaper articles. It was programmed in Java and runs under Windows NT on a Pentium P60 using an MS Access database.

3. Experimental methods

Twelve users (5 males, 7 females, age range 23–46) participated in the experiment. The information searching experience of the users, captured by a pre-study questionnaire, is summarized in Table 1. Most users were researchers or students at City University, but they had diverse backgrounds and interests and represented a wide cross-section of casual and professional users.

The study consisted of five phases, organized in the following sequence.

(1) Pre-test questionnaire to capture subject experience and demographics.
(2) System training in which the basic operations of the user interface were explained by running through a typical query and evaluation of results. The concept of similarity-based searching and the metaphors used in the thesaurus and results presentations (bullseyes) were explained. The subjects were given the opportunity to ask questions and were encouraged to try out the system facilities.
(3) Experimental task; the subjects were asked to carry out two searches as follows.
    (a) Please find articles discussing company mergers using the terms available in the thesaurus. Note that it may be necessary to explore different thesaurus terms to do this effectively.

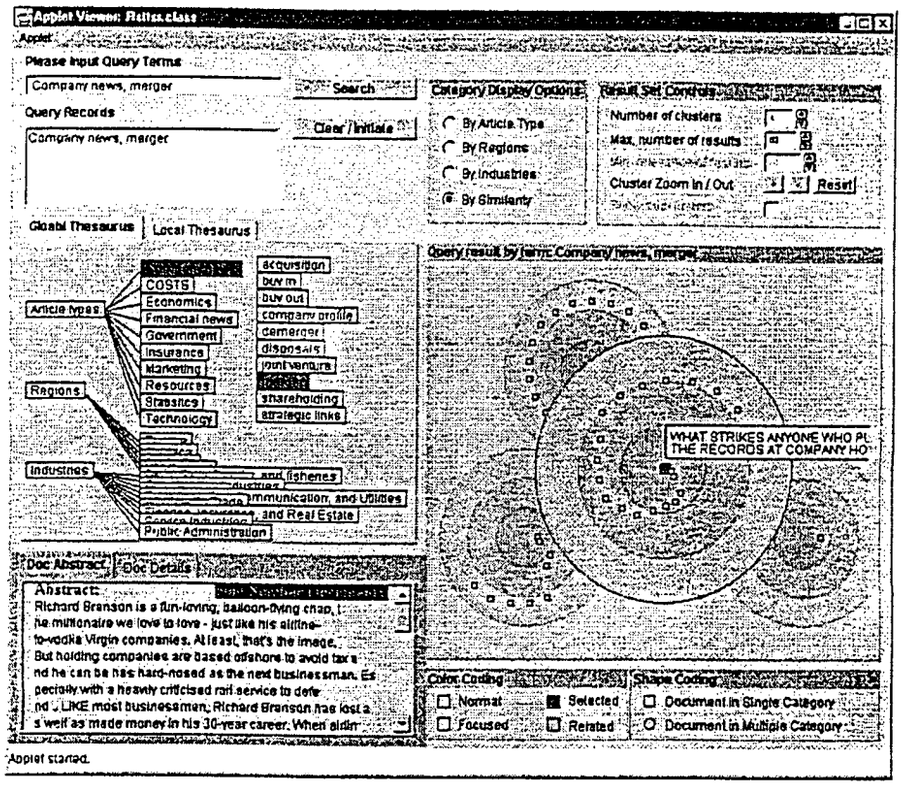


FIGURE 2. The results browser showing "hover text" titles of articles and expanded bullseye display.

TABLE 1

Pre-test questionnaire results on a cued 1–5 *scale* (< 1 *week—many times/day*)

Question	Mean score
1. Frequency of browsing for information in databases	3.1
2. Frequency of searching for information in databases	1.8
3. Frequency of browsing information on the WWW	4.8
4. Frequency of searching for information on WWW	4.0
5. Overall Web usage	4.0

(b) Please find articles which discuss the link between interest rates, inflation and wages.

Both searches were designed to encourage similarity-based searches and required multiple term queries for effective searches. The subjects were encouraged to continue searching as long as they needed to gather all the relevant articles they

thought might be present in the database. The subjects had to browse the results display, identify relevant articles from the hover text, then view the abstract and document details to determine if they were relevant. They gave the article number to the experimenter if it was judged to be relevant. All the retrieved and inspected articles were recorded. During the experiment subjects were encouraged to verbalize any problems they encountered with the system and the experimenter asked follow up questions to clarify the nature of such problems, following the practice of co-operative evaluation (Monk & Wright, 1993). The experimenter helped with usability problems only when users encountered usability breakdowns and could not proceed.

(4) After the experimental task the subjects filled in a post-test questionnaire in which they rated the usability of system facilities on a 1–5 scale.

(5) A de-briefing interview completed the session, in which the experimenter ran through a structured list of questions to probe the subjects' understanding of all the system facilities, metaphors and underlying model (i.e. similarity-based searching). In this phase, the subjects were also asked to interpret a screen diagram of the bullseye metaphor, make suggestions for improvements to the system and explain any particular problems they had experienced.

The subjects were paid £20 for their participation. Sessions lasted between 45 min and $2\frac{1}{4}$ h, with the experimental task durations ranging between 20 min and $1\frac{1}{4}$ h.

The subjects all searched the web frequently; however, database usage (bibliographic and numeric) was less frequent.

Three sets of data were collected. First, performance data for information retrieval were measured as the number of documents indicated by the subjects to be relevant to the query after viewing the abstract or the bullseye display hover text. These documents were compared with an expert's judgement of which documents in the databases were relevant, to determine a % recall measure. The precision of the subjects' retrieval was then calculated from the % of relevant documents within their retrieved set. The second dataset of usability measured the problems encountered by the subjects when using the interface as well as their reported comprehension of user-interface metaphors and functionality. The third dataset recorded subjects' behaviour patterns when interacting with the system, as a set of mental and physical behaviours.

The data were analysed to investigate individual differences and correlations between performance, usability and behaviour, to answer the following questions:

(1) Did poor usability and comprehension of the visual user interfaces result in poor performance?

(2) What patterns of usage behaviour were shown and did these correlate with performance and usability data?

4. Results

4.1. RETRIEVAL PERFORMANCE

Performance was assessed against a gold standard of relevant documents, as judged by an independent expert who was familiar with the domain and read all the articles in the test database. The maximum number of relevant documents for each task was 21 with no

TABLE 2
Recall and precision performance for both tasks by subject

	Relevant documents retrieved		Total retrieved		Recall		Precision	
	Task 1	Task 2	Task 1	Task 2	Task 1	Task 2	Task 1	Task 2
	%	%	%	%	%	%	%	%
	a	*b*	*c*	*d*	$a/21 * 100$	$b/21 * 100$	$a/c * 100$	$b/d * 100$
JU	3	4	15	23	14.3	19.0	20.0	17.4
AG1	3	4	16	18	14.3	19.0	18.8	22.2
HS	3	3	18	18	14.3	14.3	16.7	16.7
MK	2	4	13	21	9.5	19.0	15.4	19.0
KK	2	4	11	19	9.5	19.0	18.2	21.1
AG2	2	3	11	14	9.5	14.3	18.2	21.4
RM	1	3	7	13	4.8	14.3	14.3	23.1
RV	4	0	13	1	19.0	0.0	30.8	0.0
MD	2	1	10	6	9.5	4.8	20.0	16.7
AM	2	1	9	5	9.5	4.8	22.2	20.0
CN	1	1	3	4	4.8	4.8	33.3	25.0
GS	0	1	4	5	0.0	4.8	0.0	20.0
Total	25	29	130	147				
Mean	2.1	2.4	10.8	12.3	9.9	11.5	19.0	18.6

Number of relevant documents for each task = 21.

overlap between tasks, and the database contained a total of 123 documents. Recall is the number of relevant records retrieved as a percentage of the total number of relevant records in the database; precision measures the percentage of retrieved records that are relevant. The performance results are illustrated in Table 2.

Overall performance was poor. The subjects' retrieval of relevant documents ranged from 0 to 4 out of a total of 21 relevant articles in the database. Three subjects managed an average recall score $> 10\%$ for both tasks (subjects JU, AG1, HS). Overall, performance was similar on both tasks although subject RV had reasonable recall and precision in the first task but failed to find any relevant documents in the second. Most subjects selected several documents as being relevant in both tasks, with the exception of subjects CN and RV in task 2. The five worst performers (RV, MD, AM, CN and GS) also selected fewer documents, averaging 7.8 and 4.2 in the respective tasks, compared to the whole group averages of 10.8 and 12.3. Motivation may have had an effect but the subjects were well rewarded for the experiment and did take their duties seriously. Another explanation may lie in lack of domain knowledge for assessing article relevance. The database of Financial Times articles was not the subject area of any of the participants; however, the articles were not technical and an accurate assessment of relevance could have been expected by most educated members of the public.

We had no immediate explanation for the poor recall, although some reasons did emerge later when interaction with the system and usability problems were analysed.

TABLE 3
Post-test questionnaire ratings of the interface design, mean scores on 5-point semantic differential scale (e.g. very confusing ... very clear)

Question	Mean score
Presentation and structure of the thesaurus was clear	3.5
Thesaurus helped development of queries	3.4
Easy to navigate around the thesaurus	3.5
Query history was clear	2.6
System easy to use for querying	3.8
Training required for effective use	3.6
Clear relationship between query and bullseye display (results browser)	3.2
Understood the reasons for location of documents in bullseye display	4.0
Visual representation of results was clear	3.6
Bullseye display useful for navigating through results	4.1
Visual representation of results adds value over a list	4.2

4.2. SUBJECTS' RATINGS

In the post-test questionnaires the subjects rated the system favourably in spite of their poor performance. The query facilities and thesaurus (see Table 3) were rated above the mid-range on the 5-point scale, and the overall opinion on ease of querying gained a 3.8 point score. The results browsing and presentation facilities were also favourably rated by the subjects, with the bullseye display in particular receiving a 4-point plus rating. The weaker features were the history of queries (2.6) that could not be inspected easily and the link between queries and the search results display (3.2).

4.3. USABILITY EVALUATION

Usability was assessed by observing difficulties and problems encountered by users which either interrupted normal interaction or caused the user to abandon the current task. There were no occasions when users were so confused that they had to abandon the task, so the following data refer to "critical incidents" in the Monk and Wright (1993) terminology. Errors were categorized following the model mismatch evaluation method (Sutcliffe *et al.*, 2000) which classifies errors to help diagnose the underlying design feature that caused the difficulty.

The top 10 errors with their categories are given in Table 4.

TABLE 4

Usability errors observed during the tasks, categorized by design feature and problem category. Where one user experienced the same error more than once the absolute frequency is given (in parentheses)

	Error-design feature	Number of users	Error category
1	Selecting thesaurus terms	8 (12)	Execution mismatch
2	No history list for previously selected articles	7	Missing functionality
3 =	Confusing representation of the thesaurus hierarchy	6	Misleading cue + metaphor mismatch
3 =	No search progress timer	6	Missing feedback
5	Unpredictable movements in bullseye displays	5 (6)	Execution mismatch
6 =	No labels on bullseye clusters	4	Missing feedback/ functionality
6 =	Unpredictable scroll control in abstracts	4	Inappropriate functionality
8 =	Interaction not possible during search	3	Inadequate feedback
8 =	Truncated text in bullseye hover text	3	Impossible action, missing functionality
8 =	No history; repeat previous searches	3	Missing functionality

The overall frequency of errors was low. Nearly all errors were experienced only once by each user, so overall the system was usable and many of the usability problems observed pointed to missing requirements rather than design defects. The most frequent error, double clicking on the thesaurus terms to select them for a query, was caused by an excessively long inter-click interval so users found rapid double clicks had no effect. The second problem was a missing requirement for an article visit list history because users want to see the documents they had already chosen. The next (3rd =) problem, cluttered thesaurus display, obscured the hierarchical ordering; moreover, the expansion controls made links between hierarchy levels difficult to follow. In the other 3rd = problem the system did not give any feedback when the search was being executed and interaction was not possible during the search (8th = problem). Three users tried to enter another query or interact with the thesaurus map while the search was in progress, but the system only allowed single-threaded queries. The unpredictable movement of the bullseye display (6th =) was a programming error that made the display move in an unpredictable manner. The scroll control in the abstract window was also a simple programming

TABLE 5
Top ten user comprehension and usability problems reported in the post-test interviews

No.	Usability/comprehension problem	No. of users
1	Thesaurus categories and views not clear	9
2	Coding articles shared between clusters not understood	8
3	Thesaurus links between hierarchy levels not clear	7
4	No search progress indicator	7
5	Thesaurus term selection difficult	6
6	No bullseye cluster labels	5
7	Number of clusters not clear	5
8	Abstract and document details not clear	5
9	Whole display too cluttered, hard to comprehend	4
10	No history of previously selected articles	3

problem, as the scroll bar did not reset to the top of an abstract when the user moved from one document to the next. The absence of labels on the bullseye categories (6th =), although infrequent, was symptomatic of deeper problems in understanding the display metaphor. The final two (8th =) errors—truncated text in the bullseye display and no facility to reuse or repeat a previous search—point to a display feedback problem and a missing requirement.

At first sight the usability evaluation gave few reasons why user performance might be poor. The error rate was low, and most problems (see Table 5) were either missing requirements (problems 2, 4, 8 and 10), or simple programming problems (1, 5, 7 and 9), leaving only two problems that were not immediately easy to fix. The thesaurus display problem (5) was partially a consequence of the decision to use a tiled window display for consistency, thus restricting the area of the thesaurus display. However, the display clearly needed considerable improvement with better scaling/zoom controls and clearer representation of the hierarchy. The lack of labels on the bullseye clusters has no immediate solution. The LSI/cluster analysis retrieval process has no means for summarizing or identifying the relatedness of any one cluster. Human intervention is necessary to inspect the cluster of documents and assign a descriptive label summarizing the cluster's raison d'être.

The post-test de-briefing interview concentrated on further diagnosis of observed or reported problems during the session and systematically probed the users' understanding of system functionality, metaphors used in the thesaurus and results browsers and the user's model of similarity-based search. The experimenter followed a structured interview approach and asked probe questions for each area of the system in turn: thesaurus, query

formulation, results browser, abstract viewer and filter controls. Users were also encouraged to make suggestions for improving the system to remedy the problems observed during the experimental sessions. In this phase, more deep-seated problems were discovered that demonstrated that, even though superficially the system appeared to be usable, poor user understanding made usage sub-optimal. The top 10 user-comprehension problems are given in Table 5.

Several problems (1, 3, 4, 5, 6 and 10) reported by the users were identical to those observed during the test sessions, which is not surprising, although the frequency of users reporting the problem was often higher than error frequencies observed during the experimental session. More interesting are the problems not apparent in the evaluation sessions. Articles shared between two or more clusters (2), were represented as a circle whereas mono-cluster articles were shown as squares. This coding was not understood by $\frac{2}{3}$ of the users, possibly because it did not directly impinge on the experimental tasks, even though it was useful for exploring similarity and in assessing document relevance. The number of clusters in the bullseye metaphor was not clear to five subjects. The default was set at two but the users were confused by this and did not understand the system model of similarity. The number of clusters could be restricted to any arbitrary cut off the user wished; in addition, the clusters could be organized into seven industrial sector categories. Most users either stuck with the default two clusters or reduced it to one. The relationship between the abstract and the document details was criticized by five users who preferred to have the document details (author, date, keywords, etc.) first and the abstract second. Finally, four users commented that the whole display was cluttered and remarked that an overlapping rather than a tiled layout might be more effective. Two lower frequency errors not in the top 10 were lack of multi-tasking (reported by two subjects), and absence of highlighting query keywords in abstracts. The frequency of errors by subject showed a fairly even distribution, median 6, range 3–9, so only one subject encountered a high number of problems.

Debriefing interviews uncovered serious problems with the user's comprehension of the system metaphors and search functionality. The misunderstandings of metaphors and system functions that were inferred from subjects' statements in post-test interviews are given in Table 6.

Seven users experienced problems with the thesaurus. Only one user did not understand the basic hierarchy model, but she and six others had problems with the hierarchy levels and links between them, unevenness of the tree (some sub-branches had more sub-levels than others), the choice of categories and the absence of synonyms. Three users wanted to have their own customized thesaurus. Browser controls which determined the number of clusters were poorly understood, and these users also had problems with the enlarge and move functions. The concept of search by example, "find similar to this article", and use of similarity clustering as an aid to results evaluation was used by six subjects, four of whom reported that they did not understand it; a further two did not understand and did not use it. The verbal reports indicated that their model of the system, possibly influenced by experience with Web search engines (see Table 1), was a simple frequency count of keyword hits determining retrieval relevance, rather than the more sophisticated LSI similarity searching that had been explained to them. Five users confessed to not understanding the filters. Although the others said they understood the concept of filtering the retrieved result set by industry sector, date of article, etc., none of

TABLE 6
Misunderstandings of metaphors or system functionality reported by users in de-briefing interviews

Misunderstanding	No. of users
Thesaurus structure and links	7
Browser controls	6
Similarity model	6
Filters	5
Bullseye results-browser metaphor	3
Query formulation	3
Abstract viewer	1

them actually used these functions. The encoding of relevance in the browser-bullseye metaphor was understood by most users, although the rationale for clusters was not clear for three subjects. Three subjects were not aware that they could enter their own keywords as well as picking them from the thesaurus, also two subjects were confused about how to enter Boolean operators in queries. The system did not support Boolean queries because these are incompatible with LSI searching which essentially operates a conjunction (AND) style search. On the positive side all users liked the results browser display, understood the link between the "hover text" summary and selecting articles to display in the abstract viewer, as well as finding query formulation easy.

In conclusion, although the usability evaluation gave the system a reasonable assessment, it did point to some reasons for poor recall in operating the thesaurus and, more importantly, poor user comprehension of the thesaurus structure, bullseye clusters and similarity based searching. However, when performance, usability and comprehension problems were examined at the individual level, no significant correlations were found (see Table 8). Furthermore, there were no obvious associations between particular types of user problem and performance, evident in the distribution of particular comprehension and usability problems (e.g. thesaurus/bullseye display) between better and worse performing subjects. System usage was sub-optimal by most subjects. Instead of using the search facilities effectively, as the behaviour analysis in the next section shows, a majority of the subjects browsed through the abstracts rather than using similarity searching or filtering results sets. The system metaphors fared reasonably well. The thesaurus suffered some design problems but the basic hierarchy of terms was accepted, and the bullseye metaphor for relevance encoding was easy to understand. Other reasons for poor performance, such as in sub-optimal behaviour and search strategies, were investigated by the behaviour analysis reported in the following section.

5. User Behaviour

Mental and physical user behaviour during the evaluation sessions were video taped. The video tapes were analysed by viewing the tapes to create a sequential record of behaviour categories using the following definitions:

Physical behaviours (observed)	
Type/edit query	data entry/edit text in query string area
Thesaurus query	double click on thesaurus term to enter it as a query
Execute search	send query to the database
Navigate thesaurus	using thesaurus controls to expand/contract the hierarchy display
Manipulate results	altering the number of results clusters, expand or move cluster
Change cluster categories	select clusters by industry or similarity
Navigate results	browse bullseye display inspecting "hover text" article summaries
Evaluate content	scan article abstract, scroll through article and document details
Find similar	use find similar articles control

Mental behaviours (verbalized or observed)	
Read article	read article details or abstract
Select terms	decide which term to use for a query
Select document	decide whether to select or reject a document.

Sequences were divided into segments and categorized with approximate timings to the nearest minute. The video tapes were replayed when sequences of behaviour were rapid and difficult to analyse. Two independent observers analysed selected segments of the sessions and their categorizations were compared. The initial inter-observer agreement was 77%. Differences were reconciled and common coding of behaviour categories was agreed. The frequency of behaviour by subject for both tasks is shown in Table 7.

Four behaviours—selecting relevant articles; evaluating abstracts; navigating results bullseye display by inspecting articles' titles; and reading abstracts—accounted for 85% of all the behaviour. Subjects were divided into better performers who achieved $> 10\%$ recall on merged data for both tasks and $> 5\%$ recall on each task, and worse performers. All of these behaviours, except reading the abstract and total behaviour frequency, were more frequent in the first six better performing subjects (JU to AG2 vs. RM to GS, $p = 0.05$ Binomial test comparing normalized scores for each group). Manipulation of the results (changing size or moving the bullseyes) and similarity searching showed considerable individual differences. To analyse patterns and search strategies, users' behaviours were scored in sequential order and cast into a matrix so that the frequencies of transitions between behaviours could be investigated (i.e. the number of times behaviour A was followed by behaviour B, A was followed by C, etc). The matrix was converted into a behaviour network diagram to illustrate the pattern of

TABLE 7

Frequencies of behaviours for each subject observed during both experimental tasks. Two behaviours, change cluster categories and select terms, occurred very rarely and were eliminated from the following analysis

Subject	Input query	Thesaurus query	Execute search	Navigate thesaurus	Manipulate results	Navigate results	Evaluate content	Find similar	Read article	Select document	Total
JU	1	5	4	2	0	79	77	0	0	40	208
AG1	3	15	5	5	10	62	56	2	17	38	214
HS	4	8	7	4	11	70	61	0	50	32	247
MK	5	14	6	6	7	51	36	4	7	34	170
KK	1	2	2	4	0	48	40	1	27	29	154
AG2	2	3	3	4	6	83	75	5	26	34	241
RM	1	3	3	2	1	36	28	4	4	19	101
RV	5	0	6	5	6	33	30	1	11	15	112
MD	3	4	2	3	5	25	23	0	11	16	92
AM	1	2	2	2	0	14	14	0	11	13	59
CN	0	2	2	2	0	8	8	0	8	6	36
GS	3	2	6	2	6	22	17	0	14	8	81
Mean	2.4	5	4	3.4	4.3	42.2	38.8	1.1	15.5	23.7	142.9

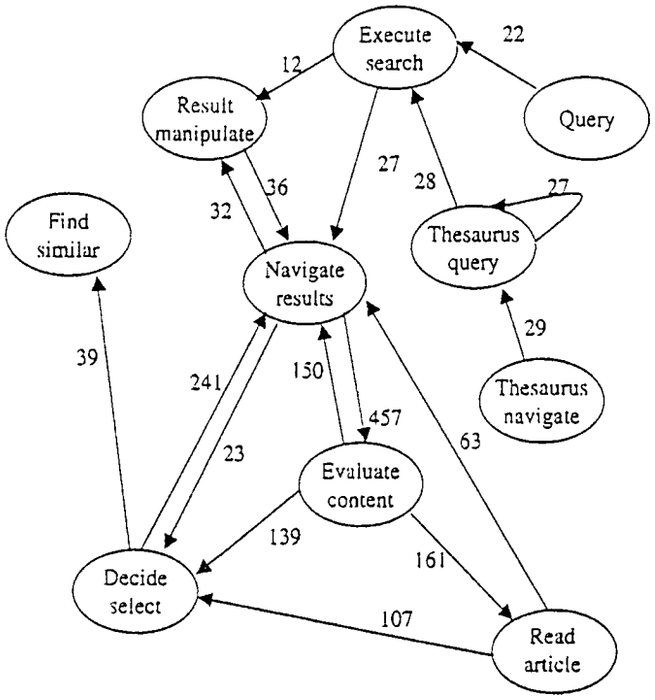


FIGURE 3. Behaviour pattern diagram for all subjects, illustrating only transitions > 1% of the total number of transitions.

behaviour during each session. The behaviour pattern for all subjects is illustrated in Figure 3.

The group-level pattern commenced with thesaurus navigation followed by selecting terms. Terms were directly entered as queries nearly as frequently as being selected from the thesaurus. Term selection led to query execution which took some time; consequently there were some breaks in the behaviour sequence at this point. At first users tried to continue interaction but they quickly learned that the system did not support multi-tasking and patiently waited until the results were displayed. The most frequent pattern was then to browse through the bullseye results display, although some users manipulated the display (usually enlarging and moving the circles) before browsing. The evaluation cycle, which started with browsing the bullseye display, was followed by evaluating the abstract by scanning, reading it through and then deciding whether to select or reject it as relevant to the task. This concluded the task in most sequences; however, in a minority of sequences the users proceeded to use the similarity search function. This was followed by more article evaluation or search termination, but it is not shown on the diagram as these transitions fall below the 1% cut off. The absence of a behaviour cycle linking the thesaurus with results evaluation or document selection, combined with the low number of queries submitted by most subjects, indicates that the users did not refine queries; instead they concentrated on evaluation of retrieved articles for a small number of queries.

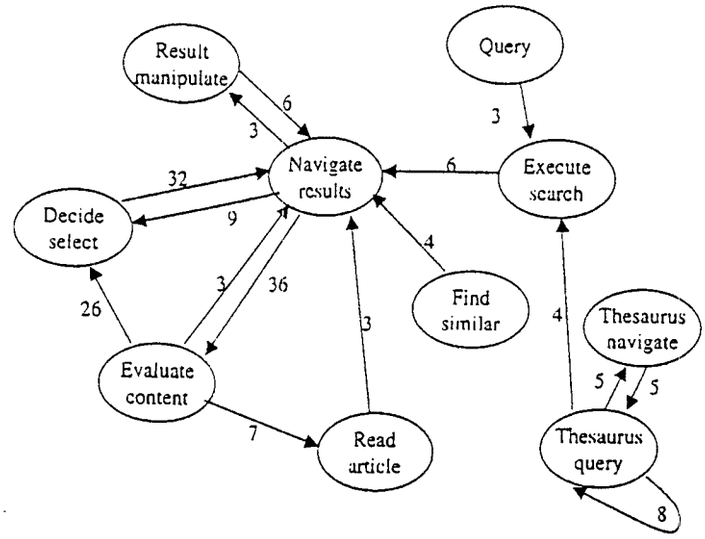


FIGURE 4. Behaviour pattern for subject MK, who was the 4th best performer in % recall. Transitions < 1% total for this subject have been omitted. This subject shows a rich behaviour pattern similar to the merged view of the whole group.

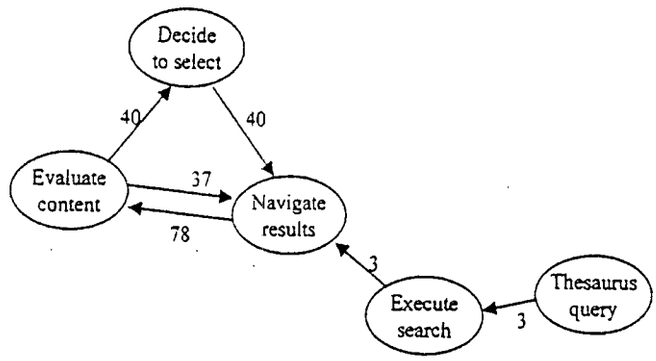


FIGURE 5. An impoverished behaviour pattern of subject JU (best performer) who did not explore the thesaurus beyond picking keywords and showed little manipulation of the results display. However, this subject did show a high frequency of transitions between Navigate results and Evaluate content, not observed for other impoverished pattern subjects.

Individual patterns either followed the group level with minor variations (see Figure 4) or showed an impoverished pattern with no manipulation of the results browser display or use of the thesaurus, as illustrated in Figure 5.

The behaviour patterns show some association with performance measures, but the picture is not consistent. Five out of the six top performers had rich behaviour patterns, but subject JU was the exception, achieving an average 16.7% recall for both tasks with an impoverished pattern (see Figure 5). However, JU did show a high frequency of transitions between navigating results and evaluating content, reflecting a systematic approach to assessing the relevance of articles. Three out of the six poorly performing

subjects (< 5% recall in at least one task) showed impoverished patterns (RV, AM, CN). The other three had richer patterns but their frequency of, and transitions between, navigating results and evaluating content were low (Table 8). These subjects read the article title on the bullseye display and made decisions to select or reject a document without viewing the abstract, so it appears that a successful strategy may require a richer pattern with frequent transitions between exploring and evaluating the results.

The association between performance, errors and search behaviour is summarized in Table 8. As noted earlier, there were no correlations between error frequencies and performance. This is not surprising as the error data contain missing user requirements as well as usability problems; furthermore, the users did not experience severe usability problems that prevented task completion.

There was a tendency for the six better-performing subjects to actively explore and navigate both the thesaurus and results browser visualizations, whereas the poorer performers only navigated the results, and did so less frequently than better performers, apart from the two subjects (RV and RM) who were the top two in the worst performer group. There is a slight tendency for better performers to submit more queries but this was not significant. Subject GS submitted many queries for no reward, although his performance may have been impaired by poor motivation, as he had to be encouraged by the experimenter to keep going several times during the task and he selected very few articles (total of nine for both tasks). Indeed, four of the six worst-performing subjects complained that they found assessing document relevance was difficult, so a combination of poor motivation and lack of domain knowledge may have accounted for their poor performance. Time spent on the task does show a positive association with performance, although this failed to achieve significance (Spearman rank order correlation coefficient). Longer task completion times, observed for three of the six better-performing subjects, were associated with more transitions between navigating results and evaluating retrieved articles rather than by other behaviours. In contrast, subject JU spent a short time on both tasks, but compensated for this by evaluating many articles carefully and selecting a large number, so there appear to be two explanations for good performance: longer task completion times and exploration of both thesaurus and results browser displays; and three explanations for poor performance: shorter task completion times, poor use of both the thesaurus and the results browser displays and possibly poor motivation leading to impaired evaluation of articles' relevance.

In conclusion, it appears our visualization was at least partially successful. A correlation between post-test misunderstandings and performance may have been expected but the data did not support this hunch, so the subjects could achieve good results in spite of sub-optimal system usage. The upside of this is to conclude that the design was robust, while the downside is that poor understanding may have restricted our subjects' retrieval performance. Usability problems did not seem to inhibit effective operations of the system. However, system facilities usage was sub-optimal and this may have constrained users from achieving better performances. Many of the problems we discovered with the users' poor understanding of the system model probably impaired performance in absolute terms, even if this did not appear to contribute to explanation of individual performance differences.

TABLE 8

Summary of user performance, usability problems, and search behaviour. The strategies are taken from the behaviour analysis, reporting only active navigation of the thesaurus and results visualizations. Recall is % for both tasks; time = Task completion time in h.mm. Infrequent navigation was scored when fewer than 20 transitions were observed between Navigate results and Evaluate content; other subjects scored > 40 transitions

Subject	Recall	Usability problems in tasks	Problems reported post-test	Report misunder-tandings	Searches submitted	Time	Strategies and behaviour
JU	16.7	5	4	1	4	0.18	Frequent navigate results
AG1	16.7	2	6	4	5	1.15	Navigate thes + results
HS	14.3	0	6	2	7	0.42	Navigate thes + results
MK	14.3	7	9	4	6	0.29	Navigate thes + results, similarity search
KK	14.3	8	7	0	2	0.34	Navigate thes + results
AG2	11.9	5	6	1	3	1.05	Navigate thes + results, similarity search
RM	9.5	7	5	6	3	0.22	Navigate results
RV	9.5	6	5	4	6	0.40	Navigate thes + results
MD	7.1	8	6	2	2	0.23	Infrequent navigate results
AM	7.1	1	4	0	2	0.21	Infrequent navigate results
CN	4.8	6	6	3	2	0.40	Infrequent navigate results
GS	2.4	7	4	4	6	0.26	Infrequent navigate results

6. Discussion

Although our visualization design appeared to pass the usability test apart from some minor glitches, user performance was poor. One reason for this may have been the relative unfamiliarity of the system. When users did interact with the visualizations, better results were achieved, so one lesson may be that more training is required. The post-test debriefing showed that the subjects had only partially understood the system metaphors and functionality, e.g. the clusters and filters. Furthermore, most of our subjects used the system with conservative strategies consisting of simple queries without cycles of refining searches. This contrasts with search behaviour that we, and others, have observed with traditional information retrieval interfaces (e.g. MEDLINE with WinSpirs) in which expert users refine queries using narrowing and broadening strategies (Marchionini, 1995; Sutcliffe, Ennis & Watkinson, in press). It is possible that visual browser interfaces inhibit such behaviour. As information search tools will often be used by end users with little training, for instance in WWW applications, expert assistants (i.e. wizards and guided tours) may be necessary to explain more complicated visualizations and how to use them effectively with efficient information searching strategies.

The implications for visualization designers from our study are that a more systematic approach to developing appropriate combinations of visualization and functionality is needed. Our design was motivated from a task model of information searching and a data model for the thesaurus and results browser. The visualization did appear to be comprehensible to users; however, it was hindered by lack of guidance on search strategies and possibly by the manipulations we provided for exploring the thesaurus and results browser visualizations. Basing visualization design on user tasks and data models has been advocated by others (Card *et al.*, 1999) and demonstrated in successful products (Ahlberg & Shneiderman, 1994); however, in more complex tasks further research on visualization design methods that integrate active system guidance with visual browser and exploration tools is required.

Our previous studies on information retrieval showed that user strategies were one important determinant of search success (Sutcliffe *et al.*, in press) as have other studies (Kulthau, 1993; Marchionini, 1995). Another success factor noted in several studies is choice of appropriate search terms (Marchionini, 1995; Ingwersen, 1996; Sutcliffe *et al.*, in press). We provided a visual thesaurus to tackle this problem; however, several users commented that the terms in the thesaurus did not match their expectations and typed in their own queries. One lesson here is that visual structure is no substitute for either a well-designed thesaurus or user-customizable thesauri. We had included a user customization facility in our design but did not test it because of increasing the complexity of an already complex system for novice users. The mismatch between terms, classification of terms and visualization structure remains a subject for future research.

The representation of results sets with the bullseye metaphor was successful, and encoding both relevance and similarity was understood by our users; hence the bullseye display appears to be an advance on simple display boxes for representing similar categories as in Scatter/gather (Pirolli *et al.*, 1996). However, closer examination of the post-test interviews showed that while the users understood the relevance ranking metaphor, their concept of functionality by which similarity was calculated by the search process was less clear. This may have been caused partly by the users' current mental

model of Web search engines overriding understanding the LSI algorithm (most users thought similarity was just shared keyword frequencies). However, LSI can occasionally rate dissimilar documents as being similar, particularly when it is being used for matching between a small number of keywords and a whole document. In our system, LSI occasionally produced some results which did not seem to belong in a particular cluster, consequently confusing the users. Another reason is probably lack of training. Our users did not make extensive use of the similarity search function although a majority stated that they did understand it.

The behaviour analysis produced a bimodal distribution between the more adventurous subjects who explored with visualization and those who were more conservative and just picked keywords and selected results. Users' prior experience did not correlate with this grouping, neither did gender, and we did not screen pre-test with a visualizer-verbalizer cognitive inventory (Leutner & Plass, 1998), so we have no explanation beyond individual difference for this effect. However, search success was not directly attributable to the users' patterns of interaction alone; instead users had to spend considerable time carefully evaluating articles as well as using the system visualization effectively. Although the bullseye results browser was approved of and used by all the subjects, it did not help the poor performers. Summarization of results can improve user performance compared to simple relevance ranking (Pirolli *et al.*, 1996), and our subjects preferred the results browser to Web search engine ranked lists; however, we found that there are considerable individual differences in people's effective use of such visualizations. Visualization tools need to encourage users to carefully inspect document contents, possibly by marking keywords in documents or hit density maps, as well as presenting overview summaries.

In our future designs we will change the tiled window screen layout, considering the complaints about the representation of the thesaurus. Also our users did not refine queries in iterative use of results browsing, thesaurus navigation and querying, so the rationale for concurrent visualization of all facilities relevant to the task was not supported. The user errors and misunderstandings suggest that a larger thesaurus display with customization facilities would be preferable. Another problem was the tendency of the visualization to bias users away from selecting their own keywords. More positively, the results browser visualization did work well with the document viewer, and these tools helped users to assess article relevance from titles on the bullseye display in combination with the document viewer. However, some of the poorly performing subjects assessed articles only from the bullseye display, so this illustrates a potential hazard of visualization tools encouraging sub-optimal and cognitive lazy practice.

The success of the visual metaphors in explaining system functionality and representation of data to users was mixed. The relevance ranking and cluster-similarity grouping metaphors were understood; however, user debriefing demonstrated that there was considerable confusion about the identity of groups of documents which the system had rated as being similar. This problem, also encountered in Scatter/gather (Pirolli *et al.*, 1996), limited the system's effectiveness because the users found it difficult to relate the clusters to their query. Similarity clustering of results may therefore not be helpful unless it is directly related to the user's query and clusters are labelled with terms related to the query. However, this requires considerable inference to analyse shared properties of all

the documents in a cluster, so most systems have to rely on manual labelling. Furthermore, the clusters have to be cohesive and consistent with the users' view of a "logical group".

Generally, our system was under-utilized. The reasons for this seem to be a combination of usability problems which impaired the effectiveness of the thesaurus, and inadequate training for more complex features such as similarity based search. In spite of under-utilization, poor performance was not directly attributable to the visualization design. Unfortunately, good visual design may be no panacea for poor search performance attributable to user motivation and lack of domain knowledge. While some studies have shown that visual information retrieval tools can improve performance over simple ranked list displays (Chen & Dumais, 2000), this study has raised a cautionary note about individual differences and demonstrates that improving users' search performance may require training and system assistance in search strategies and assessing relevance that is integrated with visual representations of meta-data and retrieved documents.

This research was partially supported by ESRC Cognitive Engineering project MISSAR (Modelling Information Seeking Strategies And Resources) and EU Telematics project Multimedia Broker.

References

AHLBERG, C. & SHNEIDERMAN B. (1994). Visual information seeking: tight coupling of dynamic query filters with starfield displays. In B. ADELSON, S. DUMAIS & J. OLSON Eds. *Celebrating Interdependence: CHI '94 conference proceedings on Human Factors in Computing Systems*, pp. 313–317. Boston. New York: ACM.

CARD, S. K., ROBERTSON, G. & MACKINLAY, J. D. (1991). The information visualizer: an information workspace. *Proceedings of CHI'91, Conference on Human Factors in Computing Systems*, pp. 181–188. New Orleans. New York: ACM.

CARD, S. K., MACKINLAY, J. D. & SHNEIDERMAN, B. (1999). Information visualization. In S. K. CARO & B. SHNEIDERMAN, Eds. *Readings in Information Visualization: Using vision to Think*, pp. 1–34. Los Altos, CA: Morgan Kaufmann.

CHEN, H. & DUMAIS S. T. (2000). Bringing order to the Web: automatically categorising search results. *Proceedings of CHI '00, Conference on Human Factors in Computing Systems*, Den Haag, April. New York: ACM.

CZERWINSKI, M., VAN DANTZICH, M., ROBERTSON, G. G. & HOFFMAN, H. (1999). The contribution of thumbnail image, mouse-over text and spatial location memory to Web page retrieval in 3D. *Proceedings of Interact '99*, Edinburgh, 1–4 September.

HANCOCK-BEAULIEU, M., FIELDHOUSE, M. & DO, T. (1995). An evaluation of interactive query expansion in an on-line library catalogue with a graphical user interface. *Journal of Documentation*, **51**, 225-243.

KULTHAU, C. (1993). *Seeking Meaning*. New York: Ablex.

INGWERSEN, P. (1996). Cognitive perspectives of information retrieval: interaction elements of a cognitive IR theory. *Journal of Documentation*, **52**, 3–50.

LANDAUER, T. K. & DUMAIS, S. T. (1997). A solution to Plato's problem: the Latent Semantic Analysis theory of acquisition, induction and representation of knowledge. *Psychological Review*, **104**, 211–240.

LEUTNER, D. & PLASS, J. L. (1998). Measuring learning styles with questionnaires versus direct observation of preferential choice behaviour in authentic learning situations: the visualizer/verbalizer behavior observation scale (VV-BOS). *Computers in Human Behavior*, **14**, 543.

MARCHIONINI, G. (1995). *Information Seeking in Electronic Environments.* Cambridge: Cambridge University Press.

MONK, A. G. & WRIGHT, P. (1993). *Improving your Human–Computer Interface: A Practical Technique.* Englewood Cliffs, NJ: Prentice Hall.

PIROLLI, P., SCHANK, P., HEARST, M. & DIEHL, C. (1996). Scatter/gather browsing communicates the topic structure of a very large text collection. *Proceedings of the ACM SIGCHI Conference on Human Factors in Computing Systems,* Vancouver. New York: ACM.

PLAISANT, C., MILASH, B., ROSE, A., WIDOFF, S. & SHNEIDERMAN, B. (1996) Lifelines, visualising personal histories. *Human Factors in Computing Systems: Proceedings of CHI 96,* Vancouver, pp. 221–227. New York: ACM.

SUTCLIFFE A. G. & ENNIS M. (1998) Towards a cognitive theory of information retrieval. *Interacting with Computers,* 10, 323–351.

SUTCLIFFE A. G, ENNIS, M. & WATKINSON, S. J. Empirical studies of end-user information searching. *Journal of the American Society for Information Science* (in press).

SUTCLIFFE, A. G. & PATEL, U. (1996). 3D or not 3D: is it nobler in the mind? In M. A. SASSE, R. J. CUNNINGHAM & R. WINDER, Eds. *People and Computers XI: Proceedings of HCI-96,* London, pp. 79–93. Berlin, Springer Verlag.

SUTCLIFFE, A. G., RYAN, M., SPRINGETT, M. V. & DOUBLEDAY, A. (2000). Model mismatch analysis: towards a deeper explanation of users' usability problems. *Behaviour and Information Technology,* 19, 43–55.

Use of Thesauri in the Full-Text Environment

Based on a paper presented at the 34th Clinic on Library Applications of Data Processing (Cochrane & Johnson, 1998)

© 1998. Jessica L. Milstead. All Rights Reserved

Introduction

The information retrieval world has changed dramatically in recent years, with the immense increase in availability of searchable full text -- and the increasing availability of powerful engines for searching the text. It is reasonable to ask whether there is any place left for thesauri in this new information retrieval environment. I believe there is a place for thesauri -- or something like them but they must change in order to continue to be of value, and it is hard to predict just what the changes will be.

A thesaurus is more than a simple equivalence list, the kind of "thesaurus" most often supported by text retrieval packages. While equivalence lists are vital to effective information retrieval, they are not enough. They can only suggest other ways of expressing an idea which is already in the user's mind; they do not remind the user of related ideas that might be valuable in searching.

A true thesaurus has equivalence relationships, but it also supports other kinds of relationships, such as genus-species, and provides navigation assistance by means of scope notes and other aids. In other words, a thesaurus is a tool designed to aid users in finding their way around a vocabulary database. In addition to its traditional use as an authority for the terms used in indexing the database, it offers reminders of terms the user might not even have considered.

The ANSI/NISO standard for thesauri (National Information Standards Organization, 1994) provides the best available information on what thesauri should do and how they should be built, but it predates the recent explosion of full text and powerful search engines, and it is not an adequate guide to future needs and potential.

The first thesauri were produced before electronic searching became widely available, but their full development coincided with the growth of online bibliographic databases. The earliest electronic files consisted only of titles, bibliographic descriptions, and indexing; if you were lucky there were abstracts, but this was not to be taken for granted when storage space was a very precious commodity and acquiring anything in electronic form generally meant rekeying it. In this environment, indexing had to be of high quality if information was to be retrieved at all. Hence the obvious need for thesauri.

Today it is beginning to seem as if all information is available in full text. However, this is not true, nor will it be true in the immediate future. Vast numbers of legacy documents remain, and converting these to searchable text is an expensive, long-term proposition. Furthermore, many documents are still being produced only in printed form. Therefore, thesauri and indexing will continue to have a place at least for awhile in facilitating access to documents for which electronic text is not available. Their long-run value, however, depends on integration with full-text search.

Thesauri And Search Engines

Thesauri actually have a place at both ends of the information access process, at both storage and retrieval. The amount of electronically accessible full text is so immense, and is growing so fast, that

users need all the help they can get in accessing it. The explosive growth of Web search engines, with their primitive algorithms, has had some rather unfortunate effects, to my mind. Some of these engines appear to have been developed by people who saw a need, but who had not the vaguest idea that there was already a history of development of tools to fulfill similar needs. There is little evidence that some of these developers had ever used either Dialog or a library catalog.

We should distinguish kinds of tools for facilitating access to full text on the basis of the attention they give to semantics. Older, exact-match (Boolean) systems give no attention to semantics. The search terms must appear in the document for it to be retrieved; if a term appears at all the document will be retrieved regardless of whether the term is important to the meaning of the document or not. Another approach relies on statistical information -- co-occurrence of words in the document, frequency, etc. Boolean and statistically-based systems have been found to have comparable retrieval performance, but to produce very different retrieval sets. That is, searches of the same database using a Boolean engine and a statistically-based one often produce about the same number of relevant hits, but there may be little overlap between the two sets of hits.

Intelligent retrieval systems integrate statistical and semantic information, as well as a full battery of linguistic techniques, to retrieve more useful results. Such a system may contain an extensive lexicon, not just of word meanings and equivalents, but of word types and relationships. Text is parsed, to a greater or lesser extent depending on the system, and there are often tools for disambiguation of terms. Phrases rather than just single words can also be handled. The most powerful systems actually can determine syntactic or structural meaning, permitting them to retrieve a concept expressed in words that are not actually in the lexicon.

Any of these types of system could produce better results by taking advantage of the presence of controlled-vocabulary indexing. In a Boolean system the chances of retrieving relevant documents that do not happen to contain the words of the search query are improved, though precision is not helped unless the search is specifically limited to controlled vocabulary terms. In either statistical or intelligent systems, the index terms could be weighted more heavily than the running text. Unfortunately, this is generally not the case.

Searchers consistently state that they need indexed, searchable full text (Pritchard-Schoch, 1993). For some kinds of queries, statistical techniques applied to the full text have been satisfactory, while others just cannot be answered satisfactorily without indexing.

Use of Thesauri in Searching

In the traditional scenario, an indexer uses the thesaurus to select index terms for inclusion in the document record. Then the searcher, hopefully referring to the same thesaurus, selects terms which seem likely to produce relevant results, and searches the indexing, retrieving on the basis of exact match. Even if the searcher has not referred to the thesaurus, s/he is aided by the indexing because if the query words appear in the indexing, then all documents indexed with those words will be retrieved, whether the words happen to appear in the text or not.

The basic design of thesauri to date, then, has been as indexing aids, with the expectation that searchers would be able to use these aids as a guide to searching. The notation used in term relationships is abstruse; the fact that "BT" and "NT" mean that two terms are related hierarchically is obvious only to specialists. Furthermore, database producers frequently do not mount their thesauri on their search systems. And even if the thesaurus is mounted, the search system may not support the full range of navigational information. In other words, the thesaurus is an indexing aid which we hope can also be used for searching, but we frequently haven't put much effort into making this use possible, let alone easy.

Thesauri are known to be underused by searchers; this is probably due at least partly to the fact that

the thesaurus for a database is unlikely to be readily available to them. Even if it is available, it may be only in paper form, or as an online list with little or no user aid in the interface. Even without significant changes in the nature of the thesaurus itself, provision of effective thesaurus navigation tools in interfaces to search systems should increase searcher use substantially. Permitting the searcher to switch seamlessly between navigating the thesaurus and searching the database can only improve access.

An obvious way in which a thesaurus can be applied directly in retrieval is to use the relationships as a means of expanding the search. Research, however, has shown that these relationships must be used with caution. In general, expanding a search to include the narrower terms tends to improve recall without great sacrifice in precision. Expanding to include broader or related terms, while it does improve recall, typically has a significant negative impact on precision.

Over the years there have been proposals for end-user thesauri designed specifically to facilitate searching. An end-user thesaurus differs from a conventional thesaurus in two primary ways: its term inclusion and organization, and its displays. It is designed to reflect and organize the total specialized vocabulary of users in a field, rather than to provide a limited list of authorized terms. It gives more information about the scope of terms, and its displays are designed around the way in which users approach information (Anderson & Rowley, 1992; Bates, 1990).

End-user thesauri have not been widely implemented, for a number of possible reasons. Conventional thesauri are costly to develop and maintain; the additional access in an end-user thesaurus would be even more costly. At the same time, until recently there seems not to have been a real understanding that the more full text there is, the more help users need in navigating it, even with a powerful search engine.

A semantic network would serve some of the same purposes as an end-user thesaurus, but would be just as costly to develop. Use of existing semantic networks such as WordNet has not shown great improvement in retrieval results, perhaps because they are not directly focused on the vocabulary of a particular user group.

Changes in Indexing

Meanwhile, as more organizations make use of machine-aided or even automatic indexing, the demands on their thesauri have increased. For many years a few organizations such as NASA, the American Petroleum Institute, and the Defense Technical Information Center have been using machine-aided indexing (MAI). In these older MAI systems, the text of titles and abstracts is run against a rule base; when a rule is matched the applicable thesaurus term is assigned to the document. The indexer reviews these candidate index terms, adding and deleting as appropriate. While their users have found that the systems increase indexer productivity significantly, there has been no great move to MAI by other database producers in the 20 or more years that these systems have been in use. Within the past few years one MAI shell system has become commercially available (Hlava & Hainebach, 1996).

The availability of powerful text analysis software could change this scenario dramatically. The same analysis used to provide relevance-ranked search results could be used to suggest candidate terms for indexing without manual development of a rule base. Already-indexed documents would be used to train the text analysis software, which would then assign candidate index terms to the documents for indexer review. Without human review, of course, the same scenario would produce automatic indexing.

To date, however, the vendors of text analysis software have concentrated their development and marketing efforts on websites and corporate intranets, with an emphasis on categorization and routing applications. Unfortunately, a system which effectively distinguishes among 100 or 1,000

categories may not scale up to 10,000 or more highly specific terms without redesign. This may be why thesaurus-based text analysis products are only now beginning to show up in the market.

MAI assumes a developed thesaurus, and ongoing maintenance and refinement of the term assignment criteria. It shifts much of the analysis effort away from review of individual documents to maintenance of the vocabulary and retraining of the system. In fact, use of MAI does not reduce the need for a thesaurus; if anything, it increases the demands made on these tools, and as a result is bringing more of their limitations to light.

Metadata and Thesauri

At first glance, it would seem self-evident that the great variety of metadata development efforts underway today would have a significant impact on thesauri. However, this is not the case so far. Metadata standards, from the Dublin Core to the CDWA (Categories for the Description of Works of Art) and beyond, are limited almost entirely to formats and frameworks, meaning standardization of the tags and of packages for them. For resource-discovery metadata, the only concern with content of the tags has been for a few limited fields such as type of resource. Metadata formats are likely to provide a means to specify the authority used for the content of the tag, but they are designed as packages and it is up to the user to design the content of the package (Milstead & Feldman, in press).

For these reasons it seems likely that metadata developments will have relatively little impact one way or the other on thesauri. Producers who are concerned with providing standardized subject access to their resources will use thesauri to determine the content of the element(s) allocated to subject metadata. Those who are less concerned about subject access will put unstandardized keywords in such metadata elements if they use them at all.

Problems of Thesaurus Design

There are fundamental problems in the basic design of thesauri that make them less than optimally useful for more powerful retrieval scenarios. There is no reason to expect that a tool designed for Boolean search on index terms will be optimized when full text is searched by a powerful engine. Unfortunately, the ways in which thesauri could be redesigned to be more useful are not immediately obvious.

The number of kinds of relationships in the present design is limited -- and yet even this specification of types is probably only of marginal direct value to many users. As indicated above, users do not necessarily recognize that "BT" and "NT" mean a relationship is hierarchical and "Use" and "UF" mean the terms are equivalent, while "RT" means the relationship is something else that something being unspecified.

For a thesaurus developer, even deciding when a relationship is hierarchical or part/whole can be difficult. The decision is fairly easy when concrete objects (e.g., truck/motor vehicle) are the issue. However, in a world where the same entity may be a "particle" (i.e., concrete) or a "wave" (not concrete), depending on how the observer happens to be looking at the entity at the moment, deciding whether something is a "thing" or a "process" may not only be difficult, it is likely to be futile.

If the distinction between hierarchical and other relationships is that porous in fact, of how much value is it to users? We find ourselves wanting to say: "These terms are very closely related, while these others, though less related, might still be useful for you," but this way of looking at relationships is not necessarily hierarchical. Instead, it involves weighting, but there is no way to build weighting into a standard thesaurus -- and the appropriate weights will be a subjective decision in any case.

At the same time, text analysis software theoretically can make use of much richer semantic analysis, not only of the relationships between terms, but of the kind of term, e.g., a process, thing, or property. Historically, this kind of analysis has been even more labor-intensive than that required to develop the relationships in a standard thesaurus. For instance, efforts such as the Cyc project have involved manual development of a knowledge base that would permit automatic analysis. On the near horizon, though, are systems which will automate development of abstractions such as relationships among concepts.

Displaying the relationships of a thesaurus in print has always involved compromises. For instance, the typical alphabetical display can only show a single level of upward and downward hierarchical relationships. Thesauri which include the full hierarchy of terms in the alphabetical display become much more voluminous. If the full hierarchical display is relegated to a separate listing, it can be difficult in the alphabetical display to show where to enter the hierarchical listing to see the full hierarchy of the term.

While electronic display of a thesaurus can ameliorate some of the limitations of the print display, making it possible, for instance, to switch back and forth between alphabetical and hierarchical display, the limitations of the screen are substituted for the limitations of the printed page. The screen display does offer possibilities of flexibility and customization that simply are not possible in print. Unfortunately, the thesauri which are currently publicly available on the Web frequently are less rich in access than the print forms of the same products.

Graphical displays have not been much used to date, but Plumb Design has implemented an interesting display of the WordNet semantic network that permits moving around and changing focus within the display.

Future of Thesauri

These tools, originally designed to facilitate consistent analysis of documents at input to an information retrieval system, are already well on their way to becoming vital retrieval tools as well. In fact, thesauri may soon be used more at retrieval than at input. They may work behind the scenes much of the time; while users should certainly have access to any available vocabulary aids if they want them, we need to design our interfaces so that users need not interact directly with the thesaurus to any greater extent than they wish or need to.

Given all the problems and limitations, how is it possible to remain positive about the need for continued use of thesauri? There are two fundamental reasons, one philosophical and one pragmatic:

- Philosophically, just as thesauri built on subject heading lists, providing more structured relationships and terms better fitted to the current searching environment, thesauri can be built on to develop vocabulary tools which meet the needs of users in the search environment of the near future.
- Pragmatically, there is increasing evidence of a realization on the part of text analysis system developers of the need to include a semantic component in their software. Whether this semantic component is a formal ANSI/NISO standard thesaurus is not as important as the fact that a rich semantic tool is embedded in the system.

A thesaurus can become the basis of a more extensive semantic network, providing information not just on what terms are used in indexing, but on how they are used within the system. Most often a semantic network includes richer relationships than a thesaurus, but there is no reason not to build the less sophisticated system, using it as a resource when it becomes feasible to develop the more powerful system.

References

Anderson, James D., and Frederick A. Rowley. "Building End-User Thesauri From Full-Text." In Barbara H. Kwasnik and Raya Fidel, eds. *Advances in Classification Research, Volume 2; Proceedings of the 2nd ASIS SIG/CR Classification Research Workshop, October 27, 1991.* Medford, NJ: Learned Information, 1992. p. 1-13.

Bates, Marcia J. "Design for a Subject Search Interface and Online Thesaurus for a Very Large Records Management Database." In: *American Society for Information Science. Annual Meeting. Proceedings, v.27.* Medford, NJ: Learned Information, 1990. p. 20-28.

Cochrane, Pauline A., and Eric H. Johnson, eds. *Visualizing Subject Access for 21st Century Information Resources; Proceedings of the 34th Annual Clinic on Library Applications of Data Processing, March 2-4,1997.* Champaign, IL: Graduate School of Library and Information Science, University of Illinois, 1998. p. 28-38.

Hlava, Marjorie M.K., and Richard Hainebach. "Machine Aided Indexing: European Parliament Study and Results." In *17th National Online Meeting. Proceedings.* Medford, NJ: Information Today, 1996. p. 137-158.

"Metadata: Cataloging by Any Other Name..." (with Susan Feldman). Online 23(1):24-31,January-February 1999. (Also available at: http://www.onlineinc.com/onlinemag/metadata/)

National Information Standards Organization. *Guidelines for the Construction, Format, and Management of Monolingual Thesauri.* Bethesda, MD: NISO Press, 1994. (ANSI/NISO Z39.19-1993)

Pritchard-Schoch, Teresa. "Natural Language Comes of Age." *Online* 17(3):33-43, May 1993.

multiplexing methods also simplify equipment. The byte-interleaved time-division multiplexing makes demultiplexing easy, because individual streams can be determined simply by position in the frame. A SONET add/drop multiplexer (ADM) is also less difficult to design and build.

BIBLIOGRAPHY

Telecommunications networking is a very active research area. The latest research results can be found in conferences such as IEEE INFOCOM, IEEE GLOBECOM, IEEE ICC, and ACM SIGCOM.

Several scientific journals publish state-of-the-art papers on the subject. These include: *IEEE Transactions on Communications, IEEE Journal on Selected Areas on Communications, IEEE/ACM Transactions on Networking, Computer Communications,* and *Computer Networks and ISDN Systems*.

Several books cover the topic in extensive detail. Most recent ones include the following.

Halsall, F., *Data Communications, Computer Networks, and Open Systems*, Addison Wesley, Reading, MA, 1996.

Schwartz, M., *Telecommunication Networks: Protocols, Modeling and Analysis*, Addison Wesley, Reading, MA, 1987.

Stallings, W., *High Speed Networks: TCP/IP and ATM Design Principles*, Prentice Hall, Englewood Cliffs, NJ, 1998.

Tanbenbaum, A. S., *Computer Networks*, Prentice Hall, Englewood Cliffs, NJ, 1996. 1996.

CHRISTOS DOULIGERIS

THESAURUS CONSTRUCTION AND NEW INFORMATION ENVIRONMENT

Introduction, or Some Remarks about Our Professional "State of the Art"

The *Encyclopedia of Library and Information Science* raises the issue of thesaurus construction not for the first time (*1*). It is no wonder that our era is seen as one of major changes in librarianship and information science. These changes were engendered by the convolution of technological progress and also led to psychological crises. The impetuous and irrepressible developments of high technology constantly make great contributions, not only in the field of information processing and retrieval, but perhaps—and even more so—to the user's and professional's whims. The former is confident that he can receive all requisite information in a twinkling of an eye, and the latter realizes that the new technological environment, which converts all past information currents into an "avalanche of information," forces him into a corner.

Our professional community has faced similar situations more than once; one can recall the appearance of computers in the quiet world of our libraries, which provoked not only turmoil but also apparent stress among the staff (*2*). As is the case with stress in general, "technostress" can become apparent in two ways: in fear of the unknown and "awesome innovations," creating at times a sense of apprehension and panic prior to approaching computers—and at times even rejecting them. Alternatively, there can be an overly passionate indiscriminate response to the discoveries of high-tech geniuses with deceptive overconfidence in their own abilities to resolve all professional problems at once. The appearance of computers in libraries immediately gave rise to the feeling that the complete automation of all library and information processes is already a *fait accompli*, not just futuristic technology. With the dawning of this innovative technology, librarians and books have become obsolete anachronisms, akin to scribes and papyrus. (Parenthetically, I must comment on the considerable contribution of science fiction to the growth and wide dissemination of this kind of illusion.) As the years pass, however, schools of librarianship continue to produce graduates as before, and their number has not diminished.

Nowadays, the Internet and World Wide Web are considered "the hits of the season." I even think we can name our time as the "Internetomania era." Once again, there are those who attempt to convince us of the ability of this worldwide information network to solve all problems, but voices of reason are not only heard, but are becoming stronger and stronger (*3*).

Life is very different from science fiction, and it is very important to understand precisely and realistically both the limits and the potential of the applications of modern technologies.

Philosophical, Psychological, and Historical Aspects

For many decades, the main opposition in the sphere of essential indexing and searching of information has been found across the languages of subject headings and classifications. Now that the era of the Internet has arrived, we once again face the same situation: retrieval of essential information from the global nets should be done by means of specially organized classification systems or subject keywords.

To begin with, we should precisely define the terms *subject indexing* and *subject searching*, their qualitative differences and distinctiveness from other kinds of semantical processing of information (classification schemes in particular), and the difference between their traditional and nontraditional images.

The fundamental distinction between subject indexing and other ways of semantical processing of information consists of its *multiaspectual* nature, which implies the following:

1. The treatment of the information, reflected during its processing, as *a single whole*
2. The expression of this information in the state of *indivisible multiaspectual* lexical units

If the first stage of general processing of subject indexing is conceptual extraction, its second stage is conceptual expression in one form or another.

At this second stage the principal methodological distinctions between principles of expression of information begin. Forms of expression can be extremely different. They are determined by constructive and functional principles of various information centers, users' contingents, and the financial possibilities of a center and its traditions.

The division of information retrieval systems into traditional or "precoordinated" (based on the principle of precoordinated indexing and searching information, found basically in the traditional card catalogs) and nontraditional or "postcoordinated" (based on the principle of postcoordinated indexing and searching information and seen first of all in the computerized systems) must be regarded as the most fundamental distinction.

"Precoordinated" are those information systems in which the coordination (or combination (*4*)) of indexing elements is rendered during information *processing*, and "postcoordinated" are those systems in which the coordination (or combination) of indexing elements is executed during information *search*.

Precoordination and postcoordination are not only two essentially different techniques of information processing and retrieval dependent on available technological possibilities, but first of all, two essentially different *philosophies* of information processing, storage, and retrieval (*5*).

If we try to unfold the correlative epistemological essence of these two methods of semantical processing of information in its historical context, we must first of all clearly understand that every society, during certain stages of its development, falls at times (sometimes frequently, even if unaware of it, on the level of social consciousness) under the great influence of what in philosophical doctrine is considered "fashionable" at that time. Such a doctrine simply enters into a certain "oscillatory resonance" with social-psychological aspirations for at least part of this society (*6*). Since information is the reflection of reality, information processing is a "reflection of a reflection," or a "reflection of a second power," and therefore it is especially subject to the influence of prevailing ideological philosophies.

The moment subject analysis made its appearance on the proscenium of American librarianship, pragmatism dominated the thinking in American society as the *advanced* philosophical doctrine, carrying on a struggle with idealism and resounding (as some researchers noted; *7*), with tones of commercialism and even anti-intellectualism. As J. H. Shera stated: "In the American character, there has been a strong strain of utilitarianism, or, rather, pragmatism, and this is perhaps no more clearly evident than in librarianship" (*8*). The emphasis on the pragmatic trend of precoordinated indexing, especially the language of subject headings, can therefore hardly come as a surprise (*9*).

Users (as in every commercial enterprise) received a specific "menu" of subject headings, which meant: "These questions are accessible to our interpretations, but others . . . Sorry!" The *indexer*, who demonstrated to users his own wording of information during its processing, played first fiddle in the theory of subject headings. The user had to trust the indexer entirely, and even though his freedom of information search was not completely limited, in a way it was exclusively dependent on the objective characteristics of an information system as well as on the subjective qualities and peculiarities (professional or personal) of an indexer (*10*).

Similarly, the appearance of postcoordination coincides simultaneously with the

rise of the movement for "freedom of information." This approach allows the user to personally choose the information, thus limiting the responsibilities of the indexer only in the extraction of *all* information from a source, yet it is the indexer who determines and organizes the efficient ways of its retrieval (*11*).

Here the personality of the indexer still plays an important role, yet he cannot be regarded as the determining factor, as it is *the user* who selects the concepts to be coordinated in the process of information searching. Postcoordination manifests itself most strikingly in the inherent bias of both conventional and nonconventional librarianship toward metaphysical interpretation of phenomena and concepts (*12*).

At the same time, one must remember that the dialectics of development imply continuity based on negation. Everything qualitatively new is born solely in the midst of that which previously existed, and from its onset it comes up against its predecessor. Just as precoordination resulted from the qualitative, revolutionary, and renewed understanding of the achievements of librarianship, postcoordination did not appear out of the blue. It was born in the depths and on the basis of the precoordinated approach, and right from the start clashed with it.

The most "common" examples of precoordinated indexing dictionaries are the subject headings' lists of dictionary and subject catalogs. One can turn to the classic Library of Congress subject headings in order to understand this concept. At that time, the ideas of C. A. Cutter about subject analysis were fundamentally novel, raising information processing and retrieval to the *qualitatively* higher level. Naturally these ideas were widely disseminated. The impact of subject headings of the Library of Congress gained momentum with the onset of two global projects—"cataloging in publication" and the system of centralized cataloging MARC (Machine-readable cataloging). Library of Congress subject headings were among the basic elements of semantic processing of information for both of these projects. According to the picturesque and rather venomous remark of S. J. Reynolds: "Subject cataloging in the United States is the province of the Library of Congress Subject Cataloging Division!" (*13*).

In due course, especially with the development of qualitatively different opportunities and such high-tech novelties as computers, the theory of subject analysis became the prevailing theme of critical overcomprehension.

The reformers of traditional subject analysis did not reject the fundamental theoretical construction of their forerunners, but only added some complementary developments in the new environment, thus the idea of postcoordination was born. It should be noted that the earliest attempts at its practical application (e.g., the use of punched cards) were, to a considerable extent, experimental. The new theories were thus in search of new technologies, which were found with the appearance of computers in the library reading halls and reference departments. As J. L. Milstead commented, now "the whole field of computer-based indexing and searching in its present form would literally be inconceivable without it" (*14*).

A crisis arose when the governing body of the Library of Congress Subject Cataloging Division was not ready for the "technological revolution." The newest electronic information carriers, while functioning within the bounds of the same systems, MARC and cataloging in publication, continued using the same old subject headings. Obviously, the precoordinated dictionaries of subject headings included in

the computerized systems acquired a certain additional flexibility, but *in fact*, they contradicted these systems. This contradiction invariably expands with time. More than two decades ago the well-known library and information specialist H. Wellisch remarked

> When we want to find out what human minds have thought and what has been preserved between the covers of a book, the MARC tapes still give us the same old LC subject headings with which we have been saddled for decades. That makes MARC rather like a modern jet plane powered by a late nineteenth-century model of a steam engine: the thing might possibly move or even fly, but it will be prone to accidents, unreliable and above all, the streamlined features of fuselage will be wasted because of the slow speed attained (*15*).

Thesauri as Unique Tools of Lexical Control and Information Retrieval

The *information retrieval thesauri* were the tools that had to resolve the contradictions. [Some questions about the principles of thesaurus construction had already been considered (*16*). Here I dwell only on the problems that seem to be the most significant and directly related to the theme of our discussion.]

There have been arguments for considering system organization as applicable to the universe as a whole, with the human language as only one of the numerous elements of its reflection. These arguments have now become mere truisms. As human language does not exist by itself, it is the specific semantic system of the reflection of objective reality. This system simulates this reality in a certain sign's form. The dictionary of this language is no other than the lexical system of conceptual reflection of all aspects of being (*17*).

Every conceptual system is subjective because

1. To a considerable extent it depends on a complex of knowledge and world outlook of its bearers (which assumes the simultaneous coexistence of a multitude of conceptual systems differently reflecting one and the same reality).
2. It is greatly exposed to changes in "space and time."

Concepts in turn crystallize to terms, which (in accordance with the property of conceptual systems just noted) can often be synonymous (i.e., one and the same concept in the different conceptual systems can be ascribed to different terms) or homonymous (i.e., different conceptual systems can fill one and the same term with essentially highly different content).

In addition, the existence of such phenomenon as terminological "prestige" must be taken into consideration. Two kinds can be described.

On the one hand, it implies the capacity of old terms to accumulate new concepts that are different, but still not crystallized. The motive of this use is very simple: very often, a used and well-known term faces considerably weaker opposition than a new one. That's why new concepts often force their way onto "trampled paths." It must be understood that in this case such "conceptual accumulators" change their traditional

connections in the term system and switch to the other structure. This process is called "transfer of meaning."

On the other hand, the process of modernization is often substituted for simple renaming. This occurrence is quite clear. One can recall how frequently of late without any basis the good old lists of subject headings are named "thesauri," a name, which *in principle* contradicts their essence, but it was said by a man of wisdom: "Into the heart of phenomenon penetrate, if rightly name it!"

Every language possesses some terminological systems embodied in its dictionary, in which specific terms of one terminological system coexist with general terms as well as with specific terms of other systems. Such discrepancies appear not only between different languages but even in the perimeters of social and national stratifications of the same language (*18*). An example is the phrase, "Cars must be kept on the pavement," which means the opposite in Great Britain as in the United States due to the antonymous meanings of the word pavement in these two English-speaking countries.

The specific conceptual subsystems with their own specific terms and terminological systems are generated within the bounds of human knowledge or activities. The terms and terminological systems, which amalgamate them, become the main material for thesaurus construction (*19*).

We can thus objectively present a coexistence of *three* different types of dictionaries, which reflect the following three lexicon levels:

1. General linguistic dictionaries, which reflect the entirety of *concepts*
2. Terminological dictionaries, which reflect the entirety of *terms*
3. Information retrieval dictionaries that are traditional and precoordinated (lists of subject headings) as well as those that are nontraditional and postcoordinated (thesauri), to either reflect the entirety of *elements of information processing and retrieval*

Thesaurus construction (as well as the construction of other information-retrieval dictionaries) is therefore the process of *simulation* of a *conceptual system* (and not of the system of natural language), which reflects objective reality or its parts on the basis of existing *terminological systems* and *objective interdependencies* between terms. Thesaurus construction as a *modeling* process does not require the existence of any *concrete* informational material for descriptor constructing at the moment of modeling; it requires only the existence of the concept itself. (Any materialization of this concept is not really important.) In thesaurus construction the *concept's existence* is the *primary* fact, whereas *the material's existence* is *secondary*. A concept as such cannot be absent in a thesaurus. It is represented there either explicitly (in the form of a functioning descriptor) or implicitly (in the form of a reference, which in the present case, besides its usual obligation of abolishing synonyms and homonyms, reserves a place in the general conceptual network for concepts, which, although unconfirmed, did not lose its authenticity; *20*). It is important to read carefully the words of one of the leading specialists in the field of scientific information, D. J. Foskett: "A thesaurus benefits greatly in use, both for indexing and for searching, if the choice of terms and relations can clearly be seen to reflect the logical structure of a subject, instead of appearing to be a haphazard collection of terms which happen to turn up in the literature" (*21*).

In its essence, simulation implies *the systems approach* to an object, and thesaurus construction is not an exception to this rule. The expediency of the application of systems analysis methods to the sphere of information science has been known for a long time, therefore I only want to mention the main points, which are relevant for our discussion.

To begin with, it is necessary to mention that "system thinking is first and foremost a point of view and a methodology arising out of this viewpoint" (*22*).

This kind of thinking—"the science of simplification" (as W. R. Ashby called systems analysis)—together with its integral part—categorial analysis—provides the opportunity

1. To approach a system that has been investigated as to its *integrated whole* with ensuing ideas about the system's *elements* as well as about its *environment*
2. To choose from the totality of system relations basic steady (*system-forming*) relations between its elements, which create a system *structure* and therefore determine its ordering, and hence, organizational degree
3. To define this structure both horizontally (concepts of the same level) and vertically (concepts of different levels), setting up its *hierarchy*
4. To carry out *system management*, the connection between different levels
5. To expose *feedback* between the system and its environment, which ensures constant system modification

As a microscope, the systems approach provides the possibility of creating a model with different degrees of approximation (from the most detailed for the central, "nuclear" concepts to the most generalized for the peripherical spheres) while simultaneously retaining the integrity of the system.

Finally, one more important factor should be taken into account. Every concept is complex in its essence and is included in a number of conceptual structures, thus retaining its integrity. To a certain extent, the image-bearing notion can provide the atom model, which is most likely quite familiar to everyone. As previously noted, a term as a stable lexical element reflects some concrete concept, which can be incorporated into various terminological systems. The same concept very often expresses differently (i.e., synonymously) in different term systems. During systems analysis of a lexical unit, and in particular, during a categorial analysis of terms according to different aspects of analysis of all of the unit, categories in different groups are provided with standard system-forming references. Amalgamated in united perimeters, all these aspects of the same term create a unique multihierarchical type—*descriptor*. It re-establishes in the multiaspectual, multihierarchical form just the complexity, which is inherent in a concept but seems to be more slack in a term. A thesaurus, which is constructed in the optimum way, must reflect the system's essence of problems, and at the same time, must be the system itself.

J. Aitchison and A. Gilchrist defined the thesaurus as "a vocabulary of a controlled indexing language, formally organized so that *a priori* relationships between concepts are made explicit" (*23*), but first, this definition puts a thesaurus in the position of an absolutely *artificial* language. (That does not entirely correspond to the facts because, as we already concluded, a thesaurus is built on the basis of a lexicon fund and semantic relationships of a whole natural language or some specific layer, and is only its model.) Second, with good reason, such a definition can also be referred to

essentially another type of information processing and retrieval means—classification schemes.

Thus, first of all, the differentiation of these two types must be clarified and their fundamental distinctions must be revealed (*24*). In the opinion of M. J. Bates, these fundamental distinctions consist of the following considerations: (1) indexing terms are grounded on a *linguistic* basis, whereas classification schemes organize *conceptual categories*; (2) the aim of indexing dictionaries is the exposure of durable solid words and expressions for document description, while the aim of classification schemes is the creation of absolutely distinct and incompatible conceptual categories that are exhaustive in their aggregate; and (3) the classification schemes' structure which is not evident in indexing dictionaries is the part of realization of such accurate division of categories (*25*). As was already noted, every information retrieval language reflects a conceptual network through the means of a lexicon, and that is why the contradictory position cited does not seem convincing. It is interesting that in the process of making his own model of information retrieval, X. Lu came to opposite conclusions from M. J. Bates's. In Lu's opinion, relationships that are merely lexical can be expressed with the structure of a classification "tree," while merely semantic relations with a structure of indexing "network" (*26*).

The fundamental difference between thesauri and classification schemes consists of their characteristic *structures* and *construction principles* and is based on the following foundations:

1. On their relation to the concept of *hierarchy*, which (following English-language explanatory dictionaries; e.g., A. S. Hornby's Oxford dictionary) can be determined as a "system with grades of authority or status from the lowest to the highest"
2. On their relation to the concept of *association*

The treatment of the concept "equivalence" by these systems is not so significant because it is present in both systems and depends first of all on the degree of development of one term's synonymic nets and its reflection in actual dictionaries.

A classification scheme initially is an indexing and retrieval tool of a *monohierarchical, monoaspectual* information system, in which every concept can be included in one or another category according to this category's aspect, whereas a priori, a thesaurus is the same type of aid for a *multihierarchical, multiaspectual* system, offering access via multiple aspects.

Multiaspectuality is the fundamental distinction of subject analysis (traditional as well as modern) from other methods for the semantic processing of information (*27*). It consists of the following:

1. The treatment of processed information as *the whole*
2. The reflection of this information by means of *indivisible* multiaspectual concepts.

For example, according to this approach "a cow" is considered as the biological object with all its anatomical and physiological features; as one of the most substantial sources of food (with its all moral and scientific pros and cons); as an object for commercial trade; as an object of religious worship of some nations; and so on. In exactly the same way a videodisc is considered here to be a certain information carrier, the physical body with its own specific properties, the object of economic relations,

and so on. As J. S. Bruner noted, "A coding system may be defined as a set of contingently related, nonspecific categories. It is a person's manner of grouping and relating information about his world, and it is constantly subject to change and reorganization" (*28*).

One and the same concept can thus *simultaneously* (according to the aspect of examination) be included into some categories of *the same* categorial system (*29*), just as this property qualitatively distinguishes categorial analysis from cluster analysis (which has now been added to the armory of librarianship and information science; *30*). These two logico-semantic methods are initially and essentially different because the categorial analysis is based on categories built *previously*, while clusters are formed *spontaneously* during analysis. It is possible theoretically and even practically to construct a separate hierarchy of relationships for every aspect of one concept or another, and the aggregation of such hierarchical "trees" must coincide with the hierarchical system of descriptors. This method was utilized in the MEDLINE thesaurus, in which a user can search an information on hierarchical trees.

The reality of this problem as well as the vital importance of multihierarchy was described many years ago by D. Soergel (*31*), but in practice we often come across the absolute misunderstanding of this thesaurus specificity, expressed in insistence on the presence of *only one* genus-related hierarchical term for every descriptor, although with intrinsic opposition some authors accept multihierarchy, but only as a rare exception to the rule. As an example, D. Batty can be cited.

> Clusters of terms should be mutually exclusive, as far as possible, that is, no term in one cluster should appear in any other clusters without good reason. Sometimes they may: the term ASPIRIN may appear in both ORGANIC CHEMICALS, as a SALICYLIC ACID, and in DEPRESSANT DRUGS. ASPIRIN, then, has two BT (Broader Term) relationships: to the higher term in ORGANIC CHEMICALS and to the higher term in DRUGS. This is known as a polyhierarchy, something that lexicographers try to avoid if possible, but accept if inevitable (*32*).

Even J. Aitchison describes multihierarchical relations ("polyhierarchy") of a descriptor only as *ones of some* relations, but not as ones *determining* thesauri's specificity (*33*). Elsewhere, however, the same author asserts the following:

> The alphabetically arranged thesaurus . . . has as its base a classificatory structure which is not immediately apparent. . . . a powerful tool for organizing the subject field is facet analysis, which engenders a more precise understanding of the terms and how they are interrelated. In facet analysis, a systematic schedule or faceted classification is constructed from which is later derived the conventional alphabetical thesaurus (*34*).

In this connection, J. Aitchison's solution to the problem of "the repetition of identical concepts in several places in the schedules" (i.e., the multiaspectual essence of a concept, and hence a term) is extremely significant. The author proposes the following way of dealing with this problem:

> "While the same thesaurus term could be linked to more than one class number, thesaurus compilation is simplified if a preferred place is selected for the concept in the schedules and a cross-reference made, in the form of hierarchical or associative relationships between the preferred and non-preferred location" (*35*).

It seems that this assertion clearly demonstrates an absolutely classification-based way of thinking because "a preferred place" can be selected only in classification schemes and not in alphabetically arranged thesauri. In addition, in this case we are dealing with a very strange interpretation (in the confines of thesaurus construction) of the term cross-reference as a form of "hierarchical or associative relationships between preferred and non-preferred location." It is obvious that Aitchison is speaking about descriptors, but she refers in this case only to the location in classification schedules because the relationships between preferred and nonpreferred terms are not hierarchical or associative but *equivalent*. There is only one conventional way of expressing these relationships between preferred descriptor and nonpreferred terms—references, which do not have a specific place in a network of relationships, even in the case of quasi-synonyms.

It seems that here we come across draw between absolutes of the classic ideas of S. Ranganathan and the misunderstanding of the real essence of systems theory, a misunderstanding that was expressed by the following assertion of H. Iyer: "The systems approach is hierarchic in nature and moves from the particular to the general and also vice versa" (*36*). In fact, Ranganathan's classification can be interpreted as *the most system-organized classification scheme of all existing hierarchical ones*, but only of them specifically.

Besides, as H. Iyer stresses, Ranganathan's method "is free from linguistic and cultural influences" (*37*), but it is impossible to eradicate a concept from its environment, which not only influences it but changes itself, and therefore changes this concept. The facet principle therefore forces us to drive every term into a Procrustean bed of constant facets, whereas

1. Not every term can be easily categorized, and this difficulty seems to grow with the growth of the level of semantical and conceptual generalization.
2. Every conceptual unit has (side by side with these global categories) its own specific ones.

It should thus be understood that the system character of the thesaurus is communicated by means of categorial analysis, and not by means of facet analysis in Ranganathan's sense. Moreover, now we have some confusion that has to be taken into account over the meaning of facet. Ranganathan and his followers understood this concept very rigidly, however, there is another meaning of facets as "characteristics of subdivision, also called 'categories' or 'types of terms'... specific for each field" (*38*). Such an interpretation is much closer to categorial analysis, formerly described. Of course, the accurate usage of terminology would be extremely useful. On the other hand, such conceptual expansion can serve as a brilliant example of the terminological prestige that was discussed above. Besides, we previously saw the substitution by Batty of the term facet by the term cluster, which seems to be incorrect because of the difference of principles between these two concepts.

Finally, the recently published analysis of L. F. Spireti (*39*) does not solve many substantial questions, which the use of facet analysis for thesaurus construction raises (in particular, the decision about synonym and homonym problem).

Undoubtedly, one primary distinction between thesaurus and classification is even more important—the role of unique *associative* relations. (The significance of this type of relation is emphasized by S. C. Biswas and F. Smith; *40*). Just their *inalienable*

inherence makes this type of relation peculiar to information retrieval dictionaries used in traditional as well as nontraditional subject analysis—lists of subject headings and thesauri. This must be allowed, along with auxiliary relationships of associative character, which are built from time to time in some classification schemes.

This combination of hierarchical and associative (in other words, vertical and horizontal) relationships gives a thesaurus a peculiar scope (multimeasurable and spatial), which forms its structure similar to that of a "semantic net," which was the object of a previous study (*41*). If the classification scheme can be compared with a tree, the thesaurus (as was already mentioned) is similar to a structure of an atom. This "atomic" resemblance results not only from specific space, but also from coexistence of steady and constantly changing relations between concepts.

At the same time, it must be emphasized that the process of exposure and establishment of hierarchical and associative relationships in a thesaurus is very subtle. Decisions must be carefully considered and completely accurate. Every mistake during the realization of this process (under conditions of a thesaurus functioning in a huge database) can lead to unfortunate search results (illustrated, e.g., in *42*).

We can therefore define *thesaurus* as *multiaspectual lexico-semantic model of a term's system reflecting a certain conceptual reality, which is used as a means of information processing and retrieval*. Such a point of view on the thesaurus provides a satisfactory interpretation of reality and determines the normal relationships between the object and subject of information processing and retrieval.

A methodologically accurately constructed thesaurus essentially provides certain significant advantages to the whole information retrieval system it supports.

First, a thesaurus constructed on the system's principle qualitatively increases the "hospitality" of the whole information retrieval system to new concepts because the process of semantic simulation is the "mirror" of the cognitive processes, in which every new concept receives in any case its own proper place in the conceptual network.

Second (and this is inseparable from the first), such a philosophy of thesaurus construction sharply reduces even the possibility of divergences of conceptual interpretation of one and the same descriptor in space (by two different indexers as well as by an indexer and a user) and in time (by one and the same indexer during different periods).

Third, such a philosophy significantly reduces the duration of information retrieval, which [as R. Fidel and L. Su (*43*) clearly demonstrated] is the most important quality of an information retrieval system for users.

The Internet and Its Searching Tools

All of the information cited above focuses on the thesaurus and its role in a traditional or "pre-Internet" information environment. What is its destiny in the different new conditions created by the birth and the precipitous evolution of the World Wide Web?

First of all, we must answer the following questions. What is the Internet? What do its *real* (but not mythologized) revolutionary aspects consist of? The answer appears to be very simple: the newest global and very impressive electronic systems of *information storage, dissemination*, and *exchange*. Notice that one aspect of integral

information processing is absent in this definition—*information searching* and *retrieval*. It must be emphasized that an engine search of course gives us a chance of free, title, and other searching possibilities (*44*), but we are now discussing *thematical* or *problems*—generally speaking *essential*—search. This absence is by no means fortuitous. Why?

All the discussions of the role of lexical control in information processing and retrieval, as well as the emphasis of the advantages of free search or search by titles, and so forth, can be accounted for to a considerable extent by just these stressful conditions, the impact of nonstop high-tech developments, the real work environment of information officers. Although practical experiments as well as theoretical investigations are constantly conducted (*45*), reaffirming at the very least the necessity (and in some essential aspects of information search, even the obvious advantages) of lexical control, what are the conclusions drawn by the adherents of free search?

First, free search *in principle* is not absolutely uncontrolled, because it needs at the very least the elimination (or at a minimum, exposure) of synonymy and homonymy, this lowest stage of lexical control. Second, in the face of constantly increasing amounts of information, searching freely is useful *solely* for the exposure of supernew concepts, which are still not reflected by the aims of lexical control; otherwise, it is often only capable of absolutely irrelevant information. As a matter of principle, free search sometimes becomes a necessary *addition* to the search via the means of lexical control (*46*).

Regarding the title search, which from time to time is lauded to the sky (of course, it does not mean that it is irrelevant to our subject search of some specific source), it can beyond any doubt help under certain conditions as the *subsidiary* tool of search limitation and specification, but (like free search) it cannot be a valuable *sole* tool of thematical search (*47*). The reason is the degree of monosemantics in the titles, which, as was proved long ago, is sufficiently high in natural and technical spheres, but very low in the social sciences and humanities. Here I only want to refer to existing literature and practice evidences of limited possibilities of subject search by titles and its auxiliary character (*48*) as well as of significantly more terminological diffusion of the language of social sciences and humanities in comparison with the language of exact and natural sciences (*49*). If the languages with considerably more detailed and precisely elaborated terminology give in to title searches in general and to the problems of the humanities specifically, the conclusions drawn by Peters and Kueth seem to be in serious doubt immediately and simply unfounded after a more thorough examination.

The same situation can be observed with the Internet. Here I am deliberately ignoring the considerable expense of an Internet search, which exceeds in cost as well as in time (and everyone knows that "time is money") the expenses incurred in a CD-ROM search, because I am sure that the technological problems will be solved rapidly enough. The question is different.

Up till now a *massed* subject search (I emphasize: a massed, large-scale search for, e.g., research or academic aims, and not any search for general information purposes) by means of CD-ROM databases has been more productive, full, and qualitatively "noise-free" than an analogous Internet one. This fact, empirically well known to professional searchers, has received some recorded corroborations (*50*).

For the present, some obvious facts on which we can operate need to be pointed out. The "information explosion" led to the transformation of the information current into a sea of information (*51*). The fundamental distinction between a current and a sea is that if a current is theoretically amenable to its control, management, and organization, all contacts with a sea are possible *exclusively* on the basis of understanding its principles of existence and elaborating its rules of *navigation*. The same applies now in librarianship and information science. If the information current formerly could be managed by means of the different classic instruments of semantic processing and information search (classification schemes as well as dictionaries of subject headings), which were built on the principles of restricting and regulating the orientation ("locking" and "sewerage") of the current itself, now the question is to determine the rules of navigation through the sea of information (*52*). These rules in principle approach the designation of focal points of information space and modes of receiving information from them as well as to switching from one database to another (*53*).

These focal points of information space are the *concepts* of the surrounding world, discussed previously. We therefore speak about the exposure of these concepts, together with their intercommunications and the establishment of methods and rules of extraction of information about them (*54*).

The attempts to solve the problem of information indexing and searching on the Internet by new methods have yet to yield the desired results (*55*). All that currently exist in Internet indexing and searching tools possess a number of substantial properties, which, on the one hand, often deliver inconceivable volumes of information to users. On the other hand, Internet searching qualitatively affects search results from the standpoint of scope and relevance. For this very reason, most researchers are forced to recommend that users not rely on any single Internet search engine but to verify the retrieval mechanisms during their network search. For instance, the capacity for checking up over 140 Internet search engines simultaneously in universal query format during the same search makes the newest metasearchers not only very powerful and almost universal Internet search tools, but also indicators of the general direction of Internet development, propelled by high-tech innovations.

Now we have two major Internet searching tools, which exist in many projects: directories (or indexes) and engines (*56*). If we try to describe their characteristic traits, we find out the following:

1. Both of them are needed in different kinds of indexing.
2. Both of them are based on different kinds of lexical control.

It is precisely these two traits that reveal their fundamental organizational distinction-directories (or indexes) gravitate toward the classification principle, whereas engines are drawn towards the subject principle. Of course, every formalization suffers from simplification, but we can see just this distinction as a determining factor.

What is it? One more attempt to put new wine into old bottles? What is this situation reminiscent of? Do you recall?

We seem to be in the same situation now as we were three decades ago, when the examination of the ability of computers created obvious confusion. As often happens in similar situations, we catch at straws of repeatedly approved tools. It is by no means

a coincidence that classification schemes (be they LC, Dewey, UDC classification, or any other) as well as different controlled precoordinated vocabularies (LC subject headings, Medical Subject Headings, etc.) are already widespread on the WWW. The appeal of old proven tools has thus not lost its sweetness or its illusions.

Automatic indexing once again becomes the hit of the season. As was the case previously, it creates illusions of quick decisions in the multitude of substantial problems that arose with the avalanche of information. Also as formerly, powerful polysemy makes it suitable first of all for information processing in those spheres of knowledge in which the level of terminological development is sufficient (*57*).

Some projects of metadata structures (Dublin Core, e.g.; *58*) seem to be interesting, but they also leave common indexing problems unsettled (synonymy, homonymy, lexical control, etc.).

Thesaurus is now also a "fashionable" term. Very often the appellation thesaurus is given to information retrieval tools that have nothing in common with real thesauri, but as the saying goes, "You may many times say 'halva, halva,' but it would not be sweeter in the mouth." The use of thesauri in these new conditions indeed seems to be far more real (*59*), but thesauri also have to undergo changes. We now turn to some current decisions about thesaurus construction problems.

The idea of "thesaurus federations" is one of the most promising (*60*). This project is based on the development of "switching" rules between different thesauri included into the federation (*61*). We can now note that this trend of thesaurus construction seems to be very productive also for the linguistic provision of some global international projects, as, for example, "Sport Database," which was built and is functioning on the collaboration among a number of national sports information centers.

"Superthesaurus" construction (*62*) as well as "dynamic thesaurus" construction (*63*) also seem to be very interesting, although they are more utopian from an organizational point of view.

One more new idea that has already been realized in some projects (e.g., "Wintertree ThesDB," "MultiTes—Thesaurus Construction Made Easy," "LEXICO," Liu-Palmer's "Thesaurus Construction System," "Synaptica," "Data Harmony"; *64*) must also be discussed. It allows (as its originators maintain) the possibility of searching the Internet by aims of *your own* thesaurus. This idea is good and the systems' designs are wonderful, but it needs exactly the same careful elaboration of general thesaurus structure as well as more specific lexical problems (e.g., synonyms and homonyms), which have already been underscored. Of course we must understand that our own unsolved methodological problems of thesaurus construction will receive immediate reflection in our creation. (The examples of this appropriateness—references' disproportions, lack of terms, etc.—can be found, e.g., in the MultiTes demo thesaurus.) Such a decision gives us only the *technological* option of constructing a search tool, but we constantly have to keep in mind that *our own searching* tool will have to operate on a *strange indexing* field, therefore these thesauri's constructors must at least be very imaginative for exhaustive decisions about the aforementioned lexical problems.

Such a decision is akin to "do-it-yourself" shops, in which all elements of something can be bought and after that assembled. The initial problem of such an approach consists of the ability and skill of a "buyer". In point of fact, we see here the indirect admission of the inability to clear obstacles and turning all responsibility over to

unknown constructors of these thesauri, which would be guilty if do not cope with the task. Such a decision can be adequate for scanty nets but not to the WWW.

At the same time, it is obvious that the current systems approach to thesaurus construction may become the only productive way. In particular, it seems doubtful that decisions can be reached without such a methodological "key" during the current wide debate over such global problems as navigation on information networks to switching from one information system to another. It seems difficult, if possible at all, to construct such a specific but potentially promising "graphic" type of thesaurus, and this problem has also become a topic for discussion and has been completed in some variation on another methodological basis (*65*).

In this regard, one of the pioneers of the librarianship and information science, J. H. Shera, remarked: "Americans have always been strangely enamored of the tool, as tool, and see in it the answer to all the problems by which society is beset" (*66*). It must be acknowledged that these "growing pains" of the hyperbolical keenness for the fruits of technical progress and the blind faith in their *independent* might that were to a certain degree fed by science fiction novels, are peculiar not only to Americans, but undoubtedly are an essential element of human nature (*67*).

Life is relentless, however, and problem solving by means of magic manna will always be similar to building castles in the sky. Neither computers nor the Internet can replace a person at the stage of *semantic* processing of information; they can only *qualitatively* lighten this process. There is not, nor can there be, a valuable information search without its semantic processing, especially during its geometrical progression (*68*).

High-tech progress hardly gives us an opportunity to anticipate the outlook, which will be available to the information community in the near future, or the endless potential of technological possibilities. In any case, high technology constantly staggers us, even though we are aware of its achievements. Perhaps in the near future the search for information will be put into effect not by means of processing and regulating information currents and obtaining relevant information, but (in accordance with the ideas of "navigation") by means of global "scanning" and "locating" of these flows. Just here correctly constructed thesauri (particularly if carefully modified and with the most detailed system for establishment of synonymous relations) would be capable of bestowing a "second wind" to automatic indexing, which has yet to perform up to its full potential. To a certain extent, one can see it in the general context of the information explosion because of the inability to control terminological polysemy. In this sense the figure of speech expressed by G. M. Hodge: "Automatic support to indexing could be defined as a robotic arm used to turn the pages of an indexer's printed thesaurus" (*69*) must be considered absolutely correct.

Thesaurus and the Internet: From Virtual Illusion to World Wide Web Reality

The often impressive means of network searching that have already been mentioned can help to detect and even to decide problems of local information nets, but they will be powerless against the unruly current of the whole Internet.

The present worldwide information situation becomes more and more similar to one well-known historical project in its final stage—the Tower of Babel's construction, with its confusion of languages. Everybody knows the results of this story, and we must do everything in our power to refrain from repeating such a deplorable situation.

Facing the situation of Internet search turmoil, we must first of all take a detached view of it, which will allows us to see that we already possess some basic elements of the decision of this global problem, and that our first and maybe main task lies in the cardinal change of the *philosophy* of Internet searching media construction.

First, it is obvious that objectively uncontrolled filling of the Internet by new materials makes any attempt of traditional *controlled indexing* (in its classic sense) of absolutely impossible. What is more, this filling exists as though in isolation. Users as well as professional searchers resemble the lone fishermen; they do not know what they would catch—neither can be sure that their catch is maximal.

If we now see the Internet as an independent entity, however, we should construct a thesaurus that would also be *independent* constructed from real information. Such thesaurus must

1. Include the *maximum* number of terms and their synonyms
2. Be based on *objective* relations between terms
3. Be multilingual
4. Be "hospitable" for new terms

As I tried to demonstrate formerly, each of these features is attainable, and in my opinion, has to be integrally inherent in every modern thesaurus.

The next step is the indexing of real Internet massives. The already existing engines would make automatic indexing *on the basis of such thesaurus* and not on the basis of a subjective user's request. The "translation" of a user's question to the descriptors from the thesaurus will be done during the real search. This change can dramatically increase the level of the Internet search *relevance* and steeply reduce the information "noise" that is so disturbing now (*70*).

The construction of this "objective thesaurus" seems to be a very substantial project, yet the international information community, which already has successful experiences of similar programs (e.g., the international cooperation in the sphere of universal decimal classification under the aegis of the International Federation for Information and Documentation (FID)), also continues the efforts in this direction (which is obvious, e.g., from the activities of the working group on principles underlying subject heading languages functioning at the International Federation of Library Associations (IFLA) section on classification and indexing; *71*). If such a tested classic tool as subject headings needs international care, it is certain that Internet search problems can be decided *only* by ways of international cooperation. The growing discussion about the necessity of a common unified (that means international) Internet format is additional evidence of the problem's actuality. (*72*).

Thesauri have one noticeable and substantial "weak spot" in the eyes of standard management: their construction is relatively expensive. We must constantly bear in mind, however, that "we are not so wealthy to buy cheap things" and that "a miser pays twice."

In any case, Milstead was undoubtedly right when she claimed that thesauri in various forms have a prolonged future (*73*). In what form? Only the future will tell.

ANNOTATED REFERENCES

1. See D. J. Foskett, "Thesaurus," in *Encyclopedia of Library and Information Science*, vol. 30, A. Kent, ed. Marcel Dekker, New York, 1980, pp. 416–463; E. Svenonius, "Design of Controlled Vocabularies," in *Encyclopedia of Library and Information Science,* vol. 45, suppl. 10, A. Kent, ed. Marcel Dekker, New York, 1990, pp. 82–109; A. B. Piternick, "Vocabularies for Online Subject Searching," in *Encyclopedia of Library and Information Science*, vol. 45, suppl. 10, A. Kent, ed. Marcel Dekker, New York, 1990, pp. 399–420; J. L. Milstead, "Thesaurus Management Software," in *Encyclopedia of Library and Information Science*, vol. 51, suppl. 14, A. Kent, ed. Marcel Dekker, New York, 1993, pp. 389–407.
2. I want to note only some authors from extensive lists of works analyzing the general and concrete aspects of the concept "technostress": S. F. Fine, "Technological Innovation, Diffusion, and Resistance: An Historical Perspective." *J. Libr. Admin.*, 7(1), 83–108 (spring 1986); J. Bichteler, "Technostress in Libraries: Causes, Effects and Solutions." *Electronic Libr.*, 5(5), 282–287 (1987); S. Champion, "Technostress: Technology's Toll." *School Libr. J.*, 35(3), 48–51 (1988); M. Sievert, R. L. Albritton, P. Roper, and N. Clayton, "Investigating Computer Anxiety in an Academic Library." *Info. Tech. Libr.*, 7(3), 243–252 (1988); J. Kupersmith, "Technostress and the Reference Librarian." *RSR*, 20(2), 7–14, 50 (1992); K. Smith, "Toward the New Millennium: The Human Side of Library Automation (Revisited)." *Info. Tech. Libr.*, 12(2), 209–216 (1993); P. Carayon and S.-Y. Lim, "Stress in Automated Offices," in *Encyclopedia of Library and Information Science*, vol. 53, suppl. 16, A. Kent, ed. Marcel Dekker, New York, 1994, pp. 314–334; E. B. Winstead, "Staff Reactions to Automation." *Computers Libr.*, 14(4), 18–21 (1994); B. Quinn, "Reducing Stressful Aspects of Information Technology in Public Services." *Pub. Acc. Serv. Q.*, 1(4), 1–34 (1995); V. Bartlett, "Technostress and Librarians." *Libr. Admin. Mgt.*, 9(4), 226–230 (1995); K. Clark and S. Kalin, "Techostressed Out? Hope to Cope in the Digital Age." *Libr. J.*, 121(13), 30–32 (1996); R. M. Pose, K. Stoklosa, and S. A. Gray, "A Focus Group Approach to Assessing Technostress at the Reference Desk." *Ref. User Serv. Q.*, 37(4), 311–317 (1998).
3. See, for instance J. M. Mueller-Alexander and H. J. Seaton, "Researching Native Americans: Tips on Vocabulary, Search Strategies and Internet." *Database*, 17(2), 45–46, 48–54, 56 (1994); T. G. McFadden, "Has the CyberEmperor No Clothes?" *Indexer*, 19(2), 81–82 (1994); M.-A. Fitzgerald, "Misinformation on the Internet: Applying Evaluation Skills to Online Information." *Emergency Librar.*, 24(3), 9–14 (1997); A. Pollock and A. Hockley, "What's Wrong with Internet Searching." *D-Lib Mag.*, March, 1997, available at http://www.dlib.org/dlib/march97/bt/03pollock.html; M. Kratzert and D. Richey, "Ten Internet Myths." *College Undergrad. Libr.*, 4(2), 1–8 (1997); D. Orzech, "In Pursuit of Intelligent Search Engines." *Datamation*, May (1998), available at http://www.datamation.com/Plugin/issues/1998/may/05.search.html; W. Koehler, "An Analysis of Web Page and Web Site Constancy and Permanence." *JASIS*, 50(2), 162–180 (1999). See also N. Auer, Bibliography on evaluating Internet resources, available at http://refserver.lib.vt.edu/libinst/critTHINK.HTM.
4. D. Soergel prefers to use the term *combination* (see D. Soergel, *Indexing Languages and Thesauri: Construction and Maintenance*, Melville, Los Angeles, 1974), but I don't see any particular reason for the special demarcation of these two really synonymous words. Apropos, as we discuss terminological problems, I want to note as very strange and muddling the attempt of A. B. Piternick to introduce one more pair of terms, *precontrol* and *postcontrol* searching vocabularies, named as "analogous to the concepts of pre- and postcoordinated indexing" (A. B. Piternick, "Vocabularies for Online Subject Searching," in *Encyclopedia of Library and Information Science*, vol. 45, suppl. 10, A. Kent, ed. Marcel Dekker, New York, 1990, p. 399). Here we deal with the wrong interpretation of concepts because the lexical control as the concrete phase of information processing can be concerned with all types of indexing dictionaries, and there is only one concept opposed to a controlled vocabulary—a free lexicon! Controlled vocabularies can be *precoordinated* (e.g., subject headings' lists) and *postcoordinated* (thesauri), but this is absolutely another matter! For an accurate analysis of controlled vocabularies see E. Svenonius, "Design of Controlled Vocabularies," in *Encyclopedia of Library and Information Science*, vol. 45, suppl. 10, A. Kent, Marcel Dekker, New York, 1990, pp. 82–109.

5. The problem of precoordination and postcoordination was hitherto regarded as axiomatical and self-evident, although its philosophical as well as methodological aspects have many "reefs," therefore I am very glad to refer to an analysis of these cardinal questions which cannot at all be considered final but is nevertheless very useful: A. Kambil and D. Bodoff, "Partial Coordination. I. The Best of Pre-Coordination and Post-Coordination. II. A Preliminary evaluation and Failure Analysis." *JASIS*, 49(14), 1270–1282 (1998).
6. Interesting observations on that score can be found in S. Maricic, "Information Science as Interfaces of the Cognitive Sphere and Society." *Infor. Proc. Mgt.*, 23(1), 33–43 (1987).
7. B. McCrimmon, ed., "American Library Philosophy: An Anthology." Shoe String Press, Hamden, CT, 1975, p. IX; H. C. Wright and J. Shira, "Librarianship, and Information Science," occasional research paper no. 3, Brigham Young University, School of Library and Information Science, Provo, UT, 1988, pp. 49–51. For interesting notes about pragmatism in librarianship see also: B. Hjørland, *Information Seeking and Subject Representation: An Activity-Theoretical Approach to Information Science*, Greenwood Press, Westport, CT, 1997.
8. J. H. Shera, "Librarianship and Information Science," in F. Machlup and U. Mansfield, eds. *The Study of Information: Interdisciplinary Messages*, Wiley, New York, 1983, p. 383.
9. The interesting observations concerning the pragmatic character of the theory of subject analysis were stated by B. Hjørland (see B. Hjørland, "The concept of 'subject' in Information Science." *J. Doc.*, 48(2), 179–181 (1992). Nearly two decades ago I (being unfortunately at that time very restricted to the approach to materials, first of all to archives) stated (on the basis of a comparative analysis of the biographies of "the father of subject indexing" C. A. Cutter and "the father of pragmatism" C. S. Peirce) a hypothesis about their prolonged *personal* contacts, and as a result, Cutter's apprehension about the methodology of pragmatism, as the saying goes, "at first hand" [see G. I. Miller, *Methodological Problems of Subject Analysis (comparative analysis of the construction principles of subject headings language in the USSR and the U.S.)*, Nauka, Moscow, 1980, pp. 60–63 (in Russian)]. Not going into the verification of this hypothesis, I want only to note that the pragmatism of Cutter's ideas does not call into question by practically nobody of specialists. See, e.g., the clear description of the philosophical essence of subject analysis given by F. S. Miksa, ed., *Charles Ammi Cutter: Library Systematizer*, Libraries Unlimited, Littleton, CO, 1977, p. 28. About relations between Peirce's theory of semiotics and subject indexing see J.-E. Mai, "The Concept of Subject," *Knowledge Organization for Information Retrieval: Proceedings of the 6th International Study Conference on Classification Research*, June 16–18, 1997, University College, London, International Federation for Information and Documentation (FID), the Hague, 1997, pp. 60–67. See also S. Rice, "Attribution and Context: The Bases of Information Retrievals (and 'Meaning')," from the electronic book *Thematics: Conceptual Access to Text and Media*, 1997, available at http://gate.cruzio.com/~autospec/accesse.htm.
10. For an analysis of the influence of the personal features of an indexer on the fruits of his activities see A. Bertrand and J.-M. Cellier, "Psychological Approach to Indexing: Effects of the Operator's Expertise upon Indexing Behaviour." *J. Info. Sci.*, 21(6), 459–472 (1995).
11. The observations and reflections about philosophical and social aspects of technological progress, in particular, its correlations with the concept of freedom expressed by L. N. Morrisett, are seen as very valuable. See L. N. Morrisett, "Habits of Mind and a New Technology of Freedom." *First Monday* (peer-reviewed journal on the Internet), 1(3), (1996), available at http://www.firstmonday.dk/issues/issue3/morrisett/index.html.
12. See, e.g., M. O. Engle, "Librarianship as Calling: The Philosophy of College Librarianship." *J. Acad. Librar.*, 12(1), 30–32 (1986); R. Grover and J. Glazier, "A Conceptual Framework for Theory Building in Library and Information Science." *Libr. Info. Sci. Res.*, 8(3), 227–242 (1986).
13. S. J. Reynolds, "In Theory There Is No Solution: The Impediments to a Subject Cataloging Code." *Libr. Q.*, 59(3), 224 (1989).
14. J. L. Milstead, *Subject Access Systems: Alternatives in Design*, Academic, Orlando, FL, 1984, p. 33.
15. H. Wellisch, "Subject Retrieval in the Seventies—Methods, Problems, Prospects," In Subject Retrieval in the Seventies: New Directions: Proceedings of an International Symposium, May 14–15, 1971, H. Wellisch and T. D. Wilson, eds. Greenwood, Westport, CT, 1972, pp. 14–15. The various critiques of Library of Congress subject headings make a huge bibliography; it is quite enough to check the annual reviews of subject analysis state of the art published in *Library Resources & Technical Services*, See J. A. Younger, "Year's Work in Subject Analysis: 1981." *Libr. Resources Tech. Serv.*, 26(3),

A FRAMEWORK FOR WORK TASK BASED THESAURUS DESIGN

MARIANNE LYKKE NIELSEN
mln@db.dk

Institute of Information Studies, Royal School of Library and Information Science, Langagervej 4, DK 9220 Aalborg Ost, Denmark

Design and construction of indexing languages require thorough knowledge and understanding of the information environment. This empirical study investigated a mixed set of methods (group interviews, recollection of information needs and word association tests to collect data; content analysis and discourse analysis to analyse data) to evaluate whether these methods collected the data needed for work domain oriented thesaurus design. The findings showed that the study methods together provided the domain knowledge needed to define the role of the thesaurus and design its content and structure. The study was carried out from a person-in-situation perspective. The findings reflected the information environment and made it possible to develop a thesaurus according to the characteristics of the work domain. It seemed more difficult to capture the needs of the individual user and adapt the thesaurus to individual characteristics.

1. INTRODUCTION

The starting point for the present investigation was a real-life problem situation. In relation to a document management project a large company considered developing a company-specific thesaurus to support exchange and retrieval of documents across the organisation. No-one in the project team had a clear idea what kind of thesaurus to develop, the topical focus of the thesaurus was not defined, the sources and methods to use were not known etc. As recommended in the classical literature on thesaurus construction (Soergel, 1974; Lancaster, 1986; Aitchison et al., 1997) it was decided that a study should be carried out to investigate what kind of conceptual tool would be appropriate to secure retrieval. A thesaurus was on the agenda, but the design was to be based on the results of the domain study.

This paper presents the design and results of this empirical study. The aim is to examine what aspects or variables to investigate, what methods to use for collecting the necessary data, what methods to use for analysing data, and how to translate the results to thesaurus design. In particular, the aim is to compare different methodologies in order to provide guidelines for the development of domain stud-

Journal of Documentation, vol. 57, no. 6, November 2001, pp. 774–797

ies. The paper evaluates a mixed set of methods: group interviews, recollection of information needs and word association tests to collect data, and content analysis and discourse analysis of the collected data.

The following research questions are examined:

(1) What knowledge is needed to plan and design a thesaurus?
(2) What variables of the information environment are to be investigated in order to gain the knowledge needed?
(3) What kind of knowledge is produced by the methods under investigation?
(4) What are the implications for thesaurus design?

FThe paper is structured as follows: the case study is presented in Section 2; preliminary works are presented in Section 3; Sections 4 and 5 discuss the role of thesauri and what variables to investigate in a domain study; the research design is described in Section 6; the findings are reported in Section 7; and the discussion and conclusions are in Section 8.

2. THE CASE STUDY

The field study took place in a large product development company, working within the pharmaceutical industry. The project studied concerns the development of a document system to handle electronic medical submissions to legal authorities. The primary goal of the system is the production of submissions, but another important objective is to gather together documents about the company's products in order to facilitate exchange of information about products, test systems, techniques, methods etc. The project was initiated in 1998 and the project staff was organised in four groups, each taking care of a specific aspect: flow diagrams, implementation, word templates and publishing. Decisions concerning indexing and retrieval came under the project group dealing with flow diagrams. The thesaurus project was run as an independent project with a thesaurus manager, who had overall responsibility, and an external adviser. Decisions regarding indexing policy and thesaurus were placed in two different project groups and the scope of the thesaurus could potentially be wider than merely supporting the document management system, depending on the study results.

Some criteria were set beforehand. Different types of documents, coming from different departments, are stored in the document system: research reports, test reports, standard operating procedures (SOPs), articles and statistics. The common subject field of pharmaceuticals is treated from different angles in the documents: experimental, regulatory, clinical, non-clinical etc. A mix of novice and expert indexers index the documents according to a set policy. Thus, due to the topical and structural complexity of documents and the varied competence of the indexers, it is difficult to produce document representations of consistent quality representing all aspects and topics of the documents. As a consequence, the project staff had an idea beforehand that the thesaurus should support controlled as well as full-text searching. The objective of the domain study was to gather

information about the context of information retrieval in order to design the thesaurus according to information-searching behaviour[1].

3. EARLIER RESEARCH

Only a few papers present empirical research on how to carry out a domain study. Pejtersen (1980) was one of the first to base the development of an indexing system, in her case for fiction, on empirical field studies of users' information requests. During the 1970s 500 negotiations between users and intermediaries were recorded in public libraries. After the search, the user and the intermediary were interviewed in order to get background information about the search request and the user. The study showed how users tend to characterise book contents from a number of different perspectives and led to the development of a multidimensional classification system. The experience gained from the empirically based development of the Book House retrieval system was later developed to form a framework for work-centred evaluation and design of information systems, known colloquially as the 'onion model' (Pejtersen & Fidel, 1998). The cognitive task analysis of the work context is a top down process that begins with the analysis of the task domain and moves on to the user and involves factors like goals, task situations, division of work, social organisation, decision making, strategies, actors, resources and values.

Soergel (1985) also recommends the study of information needs as a basis for information system design. Like Pejtersen he emphasises systems that have as their general objective assistance in problem solving. Three sources of information on needs are recommended, and approaches and problems related to studying needs are discussed in detail:

- records of tasks, functions, problems and decisions;
- records of past use; and
- the users themselves who can describe their problems, their past use and their own assessment of their needs.

The importance of seeing a thesaurus and its functions in the context of the whole information system in which it is to be used is stressed, and the design of the system as a whole and its role within the institution have to be clear before one can start on the development of a thesaurus.

Hjørland (1997) operates with the concept of 'domain analysis', on which system development and improvements should be based. A typical domain analysis

1 Wilson (1999) has developed a model of information-seeking and information-searching research areas. The model distinguishes between 'information behaviour', defined as the more general field of investigation, 'information-seeking behaviour' as a subset of the field, particularly concerned with the variety of methods people employ to discover and gain access to information resources, and 'information-searching behaviour', defined as a subset of information seeking, particularly concerned with interactions between information user and computer-based information systems.

should examine the information structure of the knowledge domain[2], including the size of its literature, the distribution of the literature on various publication forms, its national/international structure, its citation patterns, cumulativity, diffusion, different paradigms, knowledge organisation, interdisciplinary exchange etc. Hjørland concentrates on the representation of scientific knowledge and thus provides a framework for analysis of scientific knowledge domains.

The present study takes place in a professional information environment. The thesaurus is going to support users who are involved in problem solving triggered by professional work tasks[3]. This situation is very similar to and related to the context underlying the frameworks of Pejtersen & Fidel (1998) and Soergel (1985), and like their frameworks the investigation has concentrated on gaining knowledge about the work domain. However, being involved in research, the company also forms part of a scientific environment, and Hjørland's (1997) framework also has relevance: the study includes investigations of scientific structures of communication and publication.

Recently, two empirical studies have been reported which used discourse analysis for the construction of indexing vocabularies. Talja et al. (1997) used the methodology to extract and classify the users' vocabulary and López-Huertas (1997) to structure the vocabulary of a thesaurus. Both used the methodology of discourse analysis to gain semantic knowledge to analyse and structure the vocabulary, and therefore have another focus compared to the present study which uses discourse analysis at a more abstract level to gain insight into the context of information retrieval.

4. PERSPECTIVE OF THE DOMAIN STUDY

During recent years traditional user studies have been criticised for being individualistic and defective in being based solely on studies of individual users' information needs and information-seeking behaviour (Hjørland, 1997; Jacob & Shaw, 1998). It is argued that the individual's internal knowledge structures are shaped through participation in socially grounded domains. Knowledge is constructed through and embedded within acting, and a domain study should study the knowledge domain itself and not the individuals acting it. Other researchers take a different approach, and for them personal and contextual variables interact

[2]Hjørland (1997) defines a knowledge domain or discourse community as a scientific, scholarly or professional domain with unique structures of communication and publication, unique types of documents, specific terminology and information structures. We use this concept to define the object of a domain study, and we agree with Ørom (2000) in proposing to use discourse community as a generic concept and to see, for example, scientific domains and culturally embedded domains as specific kinds of discourse communities.

[3]In this paper a work task is considered as the task either given to, or identified by, a worker. A task has a recognisable beginning and end, and it contains stimuli and guidelines concerning goals and/or measures to be taken. A task may have different levels of complexity and may consist of a set of subtasks (Byström & Järvelin, 1995). In the context of thesaurus construction we are primarily interested in work tasks involving information retrieval.

during the search process (Ingwersen, 1996; Vakkari, 1997; Allen, 1997). Research into social and situational influences on information needs and user behaviour has demonstrated that the situations in which people find themselves have a profound effect on their information-seeking behaviour. Research also shows that high knowledge users differ from low knowledge users, or that users with different cognitive styles interact differently with information systems (Furnas et al., 1987; Fidel, 1991; Saracevic et al., 1988; Iivonen, 1995; Yang, 1997; Bates, 1998; Ennis et al., 1999). Thus, regardless of the situation, there are also individual variables that influence how a person acts. Models developed by Ingwersen (1992) and Wilson (1999) illustrate very well how searchers are bound by different kinds of barriers. Some barriers are personal, others are role related, and others again are related to the interpretation of the specific environment in which the searcher is acting, but human interaction is always mediated by an individual's particular state of knowledge (Belkin et al., 1982a, b).

The thesaurus is a tool that helps individual users to get an understanding of the vocabulary of the collective knowledge domain. Thus, the thesaurus should reflect the characteristics of the information environment in order to present the perspective of that particular knowledge domain. However, the thesaurus should also reflect the mental models and search behaviour of the individuals acting in the domain community in order to help them to access and interact with the thesaurus. Individuals are individually shaped and bound by their particular combination of education, former work experience and other social and cultural experiences. Even though they have been formed by the domain and also try to adapt to the environment in which they are acting, they have been shaped socially and will approach the knowledge domain from a perspective which depends both on their actual situation and on their individual characteristics. Allen (1997) calls it a person-in-situation approach, and the present domain study has been designed taking that approach, trying to capture individual as well as collective information behaviour.

5. THE VARIABLES TO BE COVERED BY A DOMAIN STUDY

The overall objective of the domain study is to gain insight into the context of the thesaurus and on that basis define and design the thesaurus. Thus, the next step is to decide what aspects to investigate and what data to gather. In professional knowledge domains it is the work tasks that trigger the information needs, and both Soergel (1985) and Pejtersen & Fidel (1998) suggest investigating the work tasks and the derived information needs to get an understanding of the information use and behaviour, and to define on that basis the role, scope and viewpoint of the thesaurus. Knowledge about the work context and the overall information behaviour: topics, task situations, goals, constraints and collaborators may help the thesaurus compiler to define the topical focus of the thesaurus and make decisions regarding:

- the main function of the thesaurus;
- what subject areas to cover;
- what type of and how many concepts to include; and
- what viewpoint and understanding to represent.

The work task environment forms part of a wider context. Not only do external factors like collaborators, authorities and competitors have an influence on the way concepts and terms are used and understood, the company and its staff also belong to a scientific community laying down guidelines for the vocabulary of the subject field through research and education. Larger organisations like the one under investigation may be characterised as a 'multilanguage environment' due to the fact that employees come from different knowledge domains, have different experiences and are in charge of different work tasks (Morgan, 2000). Everyone uses terms that they understand differently depending on their individual educational background, professional experience and organisational membership. Commonly used terms, which may seem clear upon first hearing or reading, often mean one thing in one context and have a slightly different meaning in another. Morgan (2000) calls this special vocabulary 'jargon' and defines it as a special vocabulary to meet the special needs of a profession. Jargon is used to facilitate communication between specialised groups. Jargons conform to the special vocabulary used by discourse communities, groups of people who, at least in the context of a particular role, hold a recognised body of 'truth statements' in common (Olsson, 1999). The domain study should investigate domain-specific languages, professional and scientific, that influence the corporate vocabulary and try to identify the nature and importance of each.

The search behaviour and the problems that the searcher is facing will also have an influence on the function of the thesaurus. Knowledge about search behaviour may help the thesaurus compiler to define the role of the thesaurus and design the conceptual content. Considering Kuhlthau's (1993) model on the information-searching process, the main and traditional objective of a thesaurus is to support formulation of the search query. During the formulation and collection stage the user and the information system (including the thesaurus) interact most, the query is formulated and may be reformulated according to feedback from the system. The thesaurus may only play a minor role during selection, i.e. the task of identifying and selecting a topic to be investigated, and in exploration, i.e. the task of extending personal understanding and being able to form a focus for the search request. However, the kind and quantity of information needed to analyse the problem situation and develop the search request depend on the searcher's prior knowledge about the search task (Patel & Ramoni, 1997). Human perception and learning of new categories are dependent on existing knowledge and mental models, and users with insufficient knowledge about the search task may profit from the conceptual information of a thesaurus through exploration of vocabulary and relationships between terms (Bates, 1986). The domain study should investigate search routines and search problems in order to define the primary role in retrieval and make decisions about what conceptual information to include to support the searcher, i.e. definition, historical background, use and understanding, synonym and hierarchical relationship, interconnected concepts.

This analysis suggests several variables to be investigated. The work tasks, related information needs and the use of information should be investigated to choose the topical focus of concepts and develop a structure reflecting the objective and knowledge structure of the information environment. The scientific discipline(s) and the work environment should be analysed to investigate whether

the topics are approached from a certain perspective. An analysis of the language use should reveal the existence and nature of domain-specific jargons. Furthermore, remembering the nature of human mental models, the study should investigate the use of vocabulary at two levels: at a collective level connected to the knowledge domains, and at an individual level connected to the user's educational background and work experience. The type and structure of the literature may determine, for example, the amount of detail required in indexing, and the quantity of literature may also influence the need for exhaustivity and specificity. Resources available, financial as well as staff resources, are other important constraints to consider. An essential resource is sufficient time for the construction and maintenance of the thesaurus, but resources for indexing are also important factors that may influence the quality of manual indexing and thus determine the function of the thesaurus – to support indexing and/or searching.

This leaves us with the following variables:

- information environment: work tasks and derived information needs; information and seeking behaviour; discourses and jargons;
- searchers: educational and professional background; search behaviour; search problems; language use;
- subject field: topics; viewpoints; concepts; vocabularies;
- literature: type; level; quantity;
- resources for indexing and thesaurus construction: competence; time.

Aspects dealing with literature and resources primarily concern decisions on indexing policy and will not be discussed further.

6. RESEARCH DESIGN

One can consider several methods for designing a domain study. Qualitative methods are best for exploring human behaviour, and a combination of multiple research methods will reveal a more varied and valid picture of the research object (Fidel, 1993; Vakkari, 1997). In the present study[4] several analyses were chosen, qualitative as well as quantitative, for a complete domain analysis and for comparing the efficiency of creating an understanding of the domain:

(1) group interviews;
(2) content analysis and discourse analysis of user requests; and
(3) word association tests.

6.1 Group interviews

The purpose of the interviews was to gain an overall insight and understanding of information-searching behaviour in electronic information sources, and of problems in relation to the electronic search process. Oral descriptions and group discussions of information use and search were collected using the methodology

[4]The domain study is discussed in detail in Lykke Nielsen (2000).

of the qualitative research interview. A purposive sample of twenty-nine participants was chosen according to their profound knowledge of the work domain and search experience. The group interviews were based on an interview guide, structured into six categories of questions:

- participants (educational background, position, work tasks, information use);
- information needs;
- information resources;
- search behaviour;
- search problems; and
- search aids.

The interviews were taped and a short interview description was prepared. To ensure that every participant had the opportunity to come up with his or her specific problems and perspectives, they were asked to write down their answers in a questionnaire containing the open questions from the interview guide and some closed questions. From the closed answers statistics were developed giving an overview of the most used information resources, search tactics, search aids and common search problems.

6.2 Analysis of user requests

The purpose of the request analysis was to get an understanding of the context of information retrieval. The data collection consisted of a randomly selected sample of fifty user requests received by the corporate librarians in the period from August 1998 – May 2000 by internal email. The request statements (Figure 1), often very short and written in a personal and confident style, are considered as adequate descriptions of the context and form of corporate information needs. Long collaboration has led to a common understanding of the context of the underlying work tasks and related information needs between the research scientists sending the requests and the librarians.

From:	Michael Jensen
To:	Pia Jørgensen
Date:	12 January 2000
Subject:	Phanquinone

Dear Pia,
I am looking for information about a substance, which is called Phanquinone. It is not a new substance, but it has been known for some years. I want to know something about the analysis of it in serum and plasma, and I want to know something about its pharmacokinetics in animals and/or in humans.
I suggest the following keywords: phanquinone, pharmacokinetics, quantitative determination of (in plasma/serum), HPLC.

Best regards
Michael

Figure 1. *Example of email request*

Table 1. *Facet analysis of requests*

Elemental concepts	*Concept types*
(1) Phanquinone	C. Drugs and chemicals
(2) Pharmacokinetics	D. Pharmacology
(3) In animals/humans	A. Organism and test system
(4) Quantitative determination in plasma/serum	J. Analytical methods and techniques
(5) HPLC	J. Analytical methods and techniques

The data were analysed using two methodologies: quantitative content analysis categorising and relating the concepts, and qualitative discourse analysis interpreting the context of the requests.

6.2.1 Quantitative content analysis The objective of the content analysis was to gather terminological and semantic knowledge for the analysis of language use. First, the requests were divided into elemental concepts by a facet analysis. Then each concept was categorised according to a list based on the MeSH tree structure. To give an example, the request in Figure 1 was divided into five elemental concept categories or facets as shown in Table 1.

In order to investigate the relationship between concept categories and to see if some combination patterns existed, a request category matrix was constructed and the frequency of each category was calculated (Table 2). Finally, the vocabulary across the requests was compared to the vocabulary of the MeSH thesaurus to investigate the choice of terms to see if the requesters used scientific terms, slang words, abbreviations etc.

6.2.2 Qualitative discourse analysis The objective of the qualitative analysis was to investigate the users' way of understanding and relating the concepts of the work domain in order to determine whether the requests represented specific perspectives of the subject field due to work tasks, values, roles and scientific approach. The analysis focused on the context of the needs, and the main objective was to determine whether discourse communities existed within the work domain and to judge whether the differences were so strong that the thesaurus should articulate and bridge them. The theory and practice of discourse analysis[5] were used to analyse the written search requests. The requests were generally very short and did not provide much information for a proper discourse analysis. Thus, other sources were used in order to establish a better basis for

[5]A discourse is defined as a way to talk about and understand experiences and concepts within the social world. Often different discourses compete with the wish to freeze one particular meaning and understanding of the world. In traditional discourse analysis, the contrasts and conflicts of different discourses are the main study objects, including the interactive and social processes which create the different discourses and conflicts (Winther Jørgensen & Philips, 1999).

Table 2. *Frequency of concept categories*

Concept type	*Frequency in requests (%)*
C. Drugs and chemicals	80
E1. CNS diseases and disorders	32
E3. Adverse events and reactions	24
U. External relations	20
D. Pharmacology	18
J. Analytical methods	8
B. Persons	8
A. Organism and test systems	4
Q. Quality	4
E2. Other diseases	2
T. Document types	2

interpretation. The sources consulted included internal research reports, guidelines, encyclopaedias and often interviews with the librarians who handled the requests.

6.3 Word association method

A pilot test was carried out to investigate whether the method reveals domain-specific language use and special approaches to the subject field. The word association method[6] is not recommended as suitable for building up a detailed mental map (Aitchison, 1994), but the associations reveal the respondents' mental models and provide the researcher with information about the respondents' way of relating terms and their use. Commonly, the response words of a controlled, primed test will be based on the respondents' work experiences and reflect daily language and domain-specific language use (Lykke Nielsen & Ingwersen, 1999).

The test involved three researchers, who gave associations to twenty-four stimulus words, selected according to the findings from the content analysis. The stimulus words were presented both in written and spoken form. The respondents had one minute to give two word associations. They were instructed to make associations according to their work tasks and to concentrate on words related to the stimulus word from a practical perspective of use and application rather than from a formal, scientific viewpoint. Furthermore, they were instructed to consider the response words as words which may replace the stimulus words in a search query.

[6]The word association method is a quantitative method that originated in cognitive psychology and linguistics. It has been used only sparingly in information science (Reisner, 1966; Rubinoff, 1966; Pejtersen, 1991; Lykke & Skrubbeltrang, 1992). In its simplest form a series of disconnected words (stimulus words) are presented orally or in writing to the respondents, who must respond with the first word that comes to mind (response words).

The situational relevance of the associations was evaluated by the respondents. The twenty-four stimulus words elicited 153 unique response words; only six of these were judged to be irrelevant to the context of the work domain. Afterwards the response words were divided into types of relationships. As may be seen from Table 3, the response words consisted of a mix of synonyms (ST) and associatively related (RT) words: among the words categorised as synonyms some generic broader (BT) and narrower words appear. In order to analyse from what perspective the associations had been made, the response words were divided into categories of scientific viewpoints: medical, pharmaceutical and domain specific. The classification was made by the thesaurus compiler, who has a medical background.

7. FINDINGS

The following sections present the findings of the domain study. The data collected by the different study methods have been merged in order to form a picture and gain an understanding of the information environment. The results are presented in three sections: work tasks and information behaviour, discourses and jargons, and search behaviour.

7.1 Work tasks and information behaviour

The analysis made it possible to divide the future users of the thesaurus into five work task groups[7] or discourse communities, each representing an approach to the overall subject field. Table 4 gives an overview of the identified work task groups. In the table the number of information needs within each work task type is indicated in parentheses. The appearance of concept categories is also indicated. However, the numerical quantity is not interesting in itself as the fifty requests have been used primarily as a starting point for the discourse analytical investigation. The categorisation of work task groups has been developed according to

Table 3. *Response words to the stimulus 'electrocardiograms'*

Stimulus word: Electrocardiograms

Response words	*Type of relationship*	*Viewpoint*
Q-T prolongation	Related term (RT)	Work domain
Antipsychotics	Related term (RT)	Work domain
Side effects	Related term (RT)	Pharmaceutical viewpoint
ECG	Synonym (ST)	Medical viewpoint
Heart rhythm	Related term (RT)	Medical viewpoint
Heart function	Related term (RT)	Medical viewpoint
Safety pharmacology	Broader term (BT)	Work domain

[7]A work task group is formed by professionals, often from several departments in the organisation, who share the same overall missions and duties. The shared mission influences the way the members of the group use and search for information.

Table 4. *Types of work tasks and information needs*

Work task group/ discourse community:	*Characteristics of work tasks and derived information needs*				
	A priori determinability	*Repetitiveness*	*Novelty of concepts*	*Common concept categories*	*Type and structure of information need*
All discourses				Drugs and chemicals, 40	
Basic research (12): Development and test of new substances and chemicals -> composition & effect	Low degree of determinability	Low repetitiveness	Known and unknown concepts	External relations, 8 Pharmacology, 2 Diseases, 2	Conscious information needs 1 or 2 interrelated facets
Clinical and non clinical tests (6): Test of new drugs in order to document effect and safety -> reaction	Low to medium degree of determinability	Some repetitiveness	Known concepts	Analytical methods, 4 Pharmacology, 2 Test systems, 1	Conscious information needs 3 or more interrelated facets
Marketing and sales (26): Sell and promote the effect and value of a product -> benefit	Medium degree of determinability	High repetitiveness	Known concepts	Diseases, 10 AE & reactions, 11 Pharmacology, 6	Conscious information needs 3 or more interrelated facets
Competitive intelligence (5): Compare the effect and value of a product with other products -> effect & benefit	Medium degree of determinability	High repetitiveness	Known and unknown concepts	Diseases, 4 External relations, 2	Conscious information needs 1 or 2 interrelated facets
Environment and safety (1): Secure the safety of the employees ->safety	High degree of determinability	High repetitiveness	Known concepts	Working conditions, 1	Conscious and verificative information needs 1 or 2 interrelated facets

general findings about information behaviour. The framework developed by Bystrøm & Järvelin (1995) has been used to identify the work task. *A priori* determinability refers to the degree of *a priori* uncertainty about the task, process and outcome. A low degree of *a priori* determinability indicates that the result, the process and the information requirement cannot be characterised in advance. High repetitiveness indicates that the task involves actions that are repeated many times and thus may be characterised as routine tasks. The tasks are case dependent, but the process and the information requirements are well known. Novelty of concepts refers to the degree of prior knowledge about the concepts involved. The information needs have also been divided into (1) verificative needs, (2) conscious, topical needs and (3) 'muddled' needs (Ingwersen, 1992) to determine level of conceptual complexity.

The works tasks connected to 'Basic research' are largely indeterminable, not repetitive and neither the process nor the information requirement is known in advance. The basic research scientists are developing new products or processes and thus investigate new and unknown topics. However, they do not consider their information needs 'muddled' or the concepts involved unknown. They know very well what they are looking for and thus their information needs are characterised as verificative or conscious, topical needs. They prefer to make very broad searches in order to review a large set of retrieved literature and their requests/queries often consist of only one single chemical substance combined with the name of a researcher (see Table 5). The strategy of looking up a known colleague or research institution is commonly used; similar results have been reported by other studies of R&D scientists (Ellis & Haugan, 1997; Hertzum & Pejtersen, 2000).

The information needs of the test scientists ('Clinical and non-clinical tests') have a much more complex structure and consist of words belonging to several concept categories, but their work tasks are more *a priori* determinable and are more repetitive compared to those of the basic research scientists. The primary task of the test scientists is to watch test activities concerned with their own products and those of competitors. The test scientists are also responsible for the development of test methodologies. Their information requirements vary with respect to novelty but may still be characterised as conscious topical needs, representing users who want to clarify or review aspects of a known subject matter. In general they deal with the reactions to drugs, and they require information which covers several interconnected facets:

(1) the influence of a chemical or drug (pharmacology);
(2) under certain conditions (diseases, persons);
(3) in relation to certain test objects (test systems: animals, tissues, persons);
(4) verified by a certain study technique (analytical methods).

The discourses of 'Marketing and sales' and 'Competitive intelligence' are very similar to that of the 'Clinical and non-clinical tests', but they are even more determinable and repetitive. An important work task is to answer questions from psychiatrists regarding the benefit of a drug: they look for information which evaluates and demonstrates the effect and value of a certain drug type, especially

Table 5. *Request formulations*

Work task group	*Sample citation*
Basic research	'I want information about J.E. Maggio, Harvard Medical School, Boston. He has written about Beta-amyloid.'
Clinical and non-clinical testing	'... a search on chlorprothixene, pharmakokinetics, plasma/serum concentrations (in patients)'
Marketing and sales	'literature on the treatment with Clopixol and/or Fluanxol of diabetic patients'
Competitive intelligence	'a search on Citalopram, Sertraline, Paroxetine, Fluxetine, Fluvoxamine, Venlafaxine and panic disorder'
Environment and safety	'a search on CAS 55501-05-8. I am interested in name and structure.'

to ensure that no serious adverse effects are connected to the treatment. They must document the safety of different kinds of patients: pregnant women, diabetic patients, overweight patients etc., demonstrated by legally approved study reports as well as post-marketing reports on registered adverse reactions. Their needs are clear cut examples of conscious topical needs:

(1) the effect and value of a drug (pharmacology, adverse events and reactions);
(2) under certain conditions (diseases, persons);
(3) verified by a certain study technique (analytical methods, document types).

Only one example of the discourse of 'Environment and safety' concerning the local staff appeared in the data set, but the discourse analysis showed that this type of information need typically represents a specific work task of high repetitiveness. The purpose is to ensure the safety of staff working with a certain chemical. Often the needs are expressed as verificative needs using the CAS number combined with the concept of safety.

7.2 Discourses and jargons

The investigation made it clear that each work task group has its own discourse, dealing with specific concepts which may be organised in facets or concept categories. The concept category 'Drugs and chemicals' forms the core of all five discourses. Roughly, one can say that the basic research scientists deal with the composition and the effect of drugs ('pharmacology'), the test scientists, the marketing and the sales people with the benefit and reaction to drugs ('adverse reactions') in relation to specific 'diseases', proved by some 'analytical methods' using specific 'test systems'. The environment people deal with the safety of 'chemicals'. Concepts from these few categories (see also Table 4) form the local

discourses. They represent the topical focus and must be covered thoroughly in the thesaurus.

The pharmaceutical industry is very regulated with a specific set of rules and guidelines for documentation and a domain-specific vocabulary. However, this regulated vocabulary is not harmonised or accessible via a single source. A variety of vocabularies and guidelines from different source organisations must be consulted in order to get an overview of the terminology. As a consequence, the scientists use many synonyms for the same concept, and the same word is used differently by the work task groups, depending on their external contacts and counterparts. One could say that each group has its own jargon.

For the individual user it is a difficult process to become familiar with the regulated vocabulary, and at an individual level users often have a limited knowledge of the standardised vocabulary. However, language knowledge depends not only on educational and professional background, but also on the users' position. It seems that work tasks have a major influence on the degree of conceptual knowledge. The basic research scientists refer to different vocabulary problems from users in the other work task groups. The basic researchers state that they primarily have problems finding synonyms and related terms for exploring and expanding the search query, whereas the test researchers ask for help in understanding slight vocabulary differences and finding the appropriate, controlled naming of concepts. They act in a very regulated work environment and are accustomed to meeting terminological demands from regulatory authorities and to using a very precise and specific vocabulary. Compared to the basic scientists they use the vocabulary in a much more strict way. The basic research scientists do not call for thorough definitions in the same way. The staff from competitive intelligence and marketing and sales are not obliged to use the language in the same precise way, and they demand a tool for recovery of the vocabulary of the different scientific approaches: biological, chemical, medical, neurological etc. in order to collect results from external sources.

The local naming of drugs creates a special situation as the name of a drug changes several times during its lifetime. When a drug is under development it appears to have a set of numbers, each representing a specific composition. Later it gets a generic name and it ends up having various trade names depending on the market, the solution and the weight. Several competing trade names are also connected to the drug. The different naming causes problems in retrieval, because drugs are indexed differently in local as well as external databases, and because researchers seldom have a clear picture of the full life story of a drug.

7.3 Search behaviour

From the group interviews (Table 6) it became clear that a variety of internal and external information systems and services are accessed.

Regarding type of information need users state that they have a clear picture of their problem situation and derived information needs. They know very well what they are looking for and they seem to have no difficulty in formulating the information need. However, a majority of the users (21 out of 29 participants) say that they often find it difficult to delimit their information needs when formulating queries. Demands from the regulating authorities make it important to retrieve as

Table 6. *Information sources*

Resource type	*No. of active users (n = 29)*
Internal databases	27
External databases	25
Internet sources	24

much relevant literature as possible. Consequently, they try to capture new ideas and directions to follow during the search process. They express a need for cross-references to synonyms and other types of equivalent terms (Table 7) in order to delimit their search queries in such a way that they may produce high recall without getting too much noise.

Their descriptions of the search process are in line with IR models and study results, which emphasise the changing nature of information needs during the search process (Ingwersen, 1996; Bates, 1998; Wang, 1997; Vakkari, 1999; Vakkari, 2000). The findings correspond to a study by Wang (1997) focusing on users' active vocabularies across the three stages: (1) request; (2) document selection; and (3) post project. The study showed that the actual vocabulary in each later stage is substantially larger in size than the previous one, broader and deeper in hierarchy, and wider in breadth. However, the main facets of the topic are stable. It seems that facets for a topic are less dynamic than the terms used to express the topic. In a similar study by Vakkari (2000) the number of both search terms and facets increases, and the findings suggest that the users' conceptual structures representing their topics become more differentiated. The vocabulary growth consists of related terms (RT), narrower terms (NT) and synonyms (ST), reflecting a narrowing and differentiating focus and a growing mastery of terminology. Vakkari concludes that not only novices in a domain will profit from a thesaurus, but, as in the present study, the thesaurus might also support the more experienced user with ideas on how to differentiate the task and interrelate the underlying topics.

The typical searcher has some search experience and is able to use advanced search techniques, like combination of different search criteria and expansion by synonyms and spelling variants (Table 8). They also possess knowledge about search tools like thesauri and it seems that they will have no problems using the

Table 7. *Conceptual information needed*

Conceptual information	*No. of active users (n = 29)*
• Synonyms	19
• Spelling variants	16
• Scientific names	12
• Hierarchical broader terms	11
• Proper names	10
• Multilingual terms	10

Table 8. *Use of search tactics*

Search tactics	*No. of active users* (n = 29)
Identify search terms by looking up:	
• known documents	17
• retrieved records	17
• expansion lists	14
• dictionaries	11
• thesauri	8
Combine terms/search criteria	21
Include synonyms	13

thesaurus. However, they tend to employ the corporate librarians for complex computerised searching in external databases, a pattern which is known from other studies on search behaviour (Ellis & Haugan, 1997).

8. IMPLICATIONS FOR THESAURUS DESIGN

The findings show that the thesaurus has two rather different functions in retrieval. Some work tasks demand profound knowledge and understanding of the special languages (jargons) of the regulatory environment. These users are best supported by references to vocabulary standards from authorities and by thorough, lexical-like scope notes explaining semantic and terminological differences. Other work tasks demand a structure of conceptual knowledge, which supports the user with ideas on how to differentiate the task and interrelate the underlying topics. The users ask for hedges of synonyms and other equivalent variants from which they may pick alternative search terms. The references should help the searcher to broaden the search and view the topic from different, related angles in order to make comprehensive searches. The findings have resulted in a traditional thesaurus consisting of a controlled vocabulary, reflecting regulated pharmaceutical jargons, enriched by a varied set of synonyms and near synonyms.

The topical focus of the thesaurus falls within the core concept categories: drugs and chemicals, CNS diseases, adverse events and reactions, external relations, pharmacology, analytical methods, persons, organism and test systems, and quality. For these term types guidelines have been developed concerning what type of conceptual information to include. The importance of the concept categories is also reflected in the indexing policy, and the core concept categories are indexed in separate fields. The core terms will more or less cover the conceptual needs of test and sales environment, whereas it may be difficult to follow proactively the needs of the basic researchers.

The domain study revealed a strong relationship between a drug product and the diseases for which it is indicated. References to related symptoms and adverse events connected to the drug products and to related external factors (partners and competitors) seem also to be of importance. The pattern of relationship has been used to develop guidelines for the structuring of (associatively) related terms.

An important side product of the domain study was identification of guidelines and glossaries representing different 'jargons' on which to base structure and scope notes. Links to these sources are included in the thesaurus records to explain the terminological choices.

The general structure of a thesaurus record is shown in Table 9. At first sight the structure may seem rather conventional, but it includes several pieces of domain-specific information. The objective of scope notes and reference to sources has been mentioned. Beyond this some specific pointers of equivalence relationships have been included in separate fields for drug products: generic name of compounds and related trademarks (in-house and competitor). The local Lundbeck number has been chosen as the preferred controlled name for in-house products, because it is the first official name of a drug product in its life cycle. The CAS number is included in the scope note. Acronyms are also placed in a separate field as they are very commonly used within the organisation, orally and in writing. The full name has been chosen as preferred term according to the regulatory language use even though users prefer to use abbreviations in their daily communication. The field for synonyms includes a variable set of synonyms to produce comprehensive searches, but the variant forms also play an important role as lead-in terms because of the heterogeneous nature of the user group in terms of background and work tasks. Hierarchical reference to the quality systems to which some of the terms are strongly related has been included as quality control plays an important role throughout the work environment.

In Figures 2 and 3 it is shown in practice how the findings are reflected in the thesaurus design. Figure 2 shows the thesaurus entry for 'Generic drugs', a concept often confused in the information environment with 'Copycat drugs', 'Generic names' and 'Trademarks'. The scope note and the related terms try to clear up the differences as presented in Figure 2.

The life story of the drug product Lu 00-010 is shown in Figure 3. This has a generic name (GN), but it was never marketed and thus has no trade name (TM). The range of narrower Lu numbers indicates the existence of a set of variants of more or less the same substance.

Generic drugs	
UF:	Generics
RT:	Copycat drugs
	Generic names
	Trademarks
SN:	A drug not protected by a trademark. A generic drug is called by its chemical name; a manufacturer assigns a brand name. The products have the same ingredients. Standard practice and most state laws require that a generic drug be generically equivalent to its brand-name counterpart. That is, it must have the same active ingredients, strength and dosage form. (On-line Medical Dictionary, Britannica)
SC:	C0 Drugs and chemicals, general

Figure 2. *Thesaurus entry for 'Generic drugs'*

Lu 00-010	
GN:	Cis(Z)-flupentixol
BT:	Lu 00-012
NT:	Lu 05-110
	Lu 08-020
	Lu 09-024
	Lu 10-040
	Lu 10-041
	Lu 13-132
RT:	Lu 09-215
SC:	C1 Lu drugs

Figure 3. *Thesaurus entry for drug product Lu 00-010*

The thesaurus entry for the drug product Lu 00-800 is an example of an entry developed according to the findings regarding structure (Figure 4). The thesaurus user is guided to related indications (AIDS dementia complex, Alzheimer's disease, Vascular dementia, Neuropathic pain), and to marketing partners (Merz).

9. DISCUSSION AND CONCLUDING REMARKS

As a whole the domain study has provided a workable picture of the knowledge domain. Knowledge and understanding about work tasks and working conditions made it possible to shape the role and topical focus of the thesaurus. In particular, the discourse analysis made clear how to delimit the subject field and develop

Lu 00-800	
UF:	1,3-Dimethyl-5-aminoadamantane
	1-Amino-3,5-dimethyladamantane
	D 145
	Memantin
GN:	Memantine
ACR:	DMAA
BT:	Anti-Parkinson agents
	Lundbeck compounds
RT:	AIDS dementia complex
	Alzheimer's disease
	Merz
	Neuropathic pain
	Vascular dementia
SN:	An orally active, uncompetitive N-methyl-D-aspartate (NMDA) antagonist. Amantadine derivative that has some dopaminergic effects. Possible indications are Vascular dementia and Alzheimer's disease. (MESH red) CAS Number: 19982-08-2.
SC:	C1 Lu drugs

Figure 4. *Thesaurus entry for drug product Lu 00-800*

definitions and relationships reflecting the work domain. The interviews provided knowledge about search behaviour in computer-based information systems and made it possible to design the content and structure of the thesaurus. The content analysis and the word association tests mostly provided conceptual knowledge, supporting the semantic and terminological analysis when terms were collected. Guidelines regarding word form, e.g. to choose controlled forms among synonyms, were developed according to findings from the word association tests and content analysis. Many surprising relationships have been revealed by the word association tests, and this seems to be a suitable method for identifying work task relationships. Different understandings and uses of a term were also revealed by the tests.

None of the study methods provided a full picture. Each method provided distinct information about the work domain, and the methods complemented each other. The content analysis, for example, provided a good overview, and it may be a fruitful combination to begin a domain study by categorising and counting the words represented in important texts and to use the resulting structure to inform the design of more qualitative analyses giving insight and providing understanding. Interviews provided background information to interpret the findings of the discourse analyses. Looking back, it may have been better to have combined the interview analysis of search behaviour with the discourse analytical follow-up interviews into one single interview. Requests may form a natural starting point for a structured interview about search behaviour, and we see no reason not to reverse the order of the investigations. In fact, it saves some investigative costs, and a possible side product could be even more concrete descriptions of information behaviour. The group interview is still recommended for broadening the discussion.

The domain study was conducted from a person-in-situation perspective in order to develop a thesaurus which would at the same time reflect the characteristics and perspective of the particular work domain and support the individual information searcher. However, focusing on work tasks and derived information needs, using group interviews and discourse analysis, the study provided a general picture of common information behaviour. The starting point was individual routines and problems, but through discussion a general pattern was formed. Furthermore, domain-specific jargon seemed to be dominant in all kinds of materials: requests, associations, guidelines, SOPs and reports, and professional jargon could easily be identified. In return, it seemed impossible to capture the specific knowledge structures of individual users and integrate them in the thesaurus. The users made an effort to adjust to the situation, and even the very individual-oriented association method mostly reflected the jargon. Only if one were to ask specifically for synonyms, could the method be used to develop a user-oriented lead-in vocabulary.

Thus, the study made it possible to embed profound, work task related information in the thesaurus entries. In the present project the conceptual information is presented using a traditional thesaurus structure dividing the terms into synonyms, BT, NT and RT. It will require both thesaurus knowledge and preliminary conceptual knowledge to interpret and use the conceptual information appropriately. The individual must interact with the knowledge structure to get an

understanding of the vocabulary and bridge the gap between the domain-specific jargon and his or her personal vocabulary. Further research is needed and will be carried out to investigate whether the users are capable of capturing and using the conceptual knowledge of the thesaurus.

REFERENCES

Aitchison, J. (1994). *Words in the mind: an introduction to the mental lexicon.* Oxford: Blackwell.

Aitchison, J., Gilchrist, A. & Bawden, D. (1997). *Thesaurus construction and use: a practical manual.* London: Aslib.

Allen, B. (1997). Information needs: a person-in-situation approach. In: *Proceedings of an International Conference on Research in Information Needs, Seeking and Use in Different Contexts, 14–16 August, 1996, Tampere, Finland.* London: Taylor Graham, 111–122.

Bates, M.J. (1986). Subject access in online catalogs: a design model. *Journal of the American Society for Information Science*, 37(6), 357–376.

Bates, M.J. (1998). Indexing and access for digital libraries and the Internet: human, database, and domain factors. *Journal of the American Society for Information Science*, 49(13), 1185–1205.

Belkin, N.J., Oddy, R.N. & Brooks, H.M. (1982a). ASK for information retrieval: - Part I. Background and theory. *Journal of Documentation*, 38(2), 61–71.

Belkin, N.J., Oddy, R.N. & Brooks, H.M. (1982b). ASK for information retrieval: Part II. Results of a design study. *Journal of Documentation*, 38(3), 145–164.

Byström, K. & Järvelin, K. (1995). Task complexity affects information seeking and use. *Information Processing & Management*, 31(2), 191–213.

Ellis, D. & Haugan, M. (1997). Modelling the information seeking patterns of engineers and research scientists in an industrial environment. *Journal of Documentation*, 53(4), 384–403.

Ennis, A., Sutcliffe, A.G. & Watkinson, S.J. (1999). Towards a predictive model of information seeking: empirical studies of end-user searching. In: Wilson, T.D. & Allen, D.K., eds. *Exploring the contexts of information behaviour.* London: Taylor Graham, 96–115.

Fidel, R. (1991). Searchers' selection of search keys I-II-III. *Journal of the American Society for Information Science*, 42(7); 490–527.

Fidel, R. (1993). Qualitative methods in information retrieval research. *Library and Information Science Research*, 15, 219–247.

Furnas, G.W., Landauer, T.K., Gomez, L.M. & Dumais, S.T. (1987). The vocabulary problem in human-system communications. *Communications of the ACM*, 30(11), 964–971.

Hertzum, M. & Pejtersen, A.M. (2000). The information-seeking practices of engineers: searching for documents as well as for people. *Information Processing & Management*, 36, 761–778.

Hjørland, B. (1997). *Information seeking and subject representation.* London: Greenwood Press.

Iivonen, M. (1995). Consistency in the selection of search concepts and search terms. *Information Processing & Management*, 31(2), 173–190.

Ingwersen, P. (1992). *Information retrieval interaction.* London: Taylor Graham.

Ingwersen, P. (1996). Cognitive perspectives of information retrieval interaction: elements of a cognitive IR theory. *Journal of Documentation*, 52(1), 3–50.

Jacob, E.K. & Shaw, D. (1998). Sociocognitive perspectives on representation. *Annual Review of Information Science and Technology (ARIST)*, 33, 131–185.

Kuhlthau, C.C. (1993). *Seeking meaning: a process approach to library and information services.* Norwood, NJ: Ablex Publishing.

Lancaster, F.W. (1986). *Vocabulary control for information retrieval.* Arlington, Va: Information Resources Press.

López-Huertas, M.J. (1997). Thesaurus structure design: a conceptual approach for improved interaction. *Journal of Documentation*, 53(2), 139–177.

Lykke, M. & Skrubbeltrang, C. (1992). *Indekseringsstrategi og søgethesaurus for UMI-informationssystemet (Development of an indexing strategy and a searching thesaurus for the UMI system).* Aalborg: Royal School of Librarianship.

Lykke Nielsen, M. (2000). Domain analysis, an important part of thesaurus construction. Methodologies and approaches. In: Soergel, D., Srinivasan, P. & Kwasnik, B., eds. *Proceedings of the 11th ASIS&T SIG/CR Classification Research Workshop, November 12, 2000*, 9–50.

Lykke Nielsen, M. & Ingwersen, P. (1999). The word association methodology – a gateway to work-task based retrieval. In: Draper, S.W., Dunlop, M.D., Ruthven, I. & van Rijsbergen, C.J., eds. *Mira '99: Evaluating interactive information retrieval, Glasgow, UK, 14–16 April 1999.* British Computer Society, 1–11.

Morgan, J.E. (2000). Speaking the same language – the importance of having a glossary of 'commonly used terms' or 'jargon' when implementing a project management methodology and managing projects. In: *Proceedings of the Project Management Institute Annual Seminars & Symposium, September 7–17, 2000.* Houston, Texas: Project Management Institute.

Olsson, M. (1999). Discourse – a new theoretical framework for examining information behaviour in its social context. In: Wilson, T.D. & Allen, D.K., eds. *Exploring the contexts of information behaviour: proceedings of the Second International Conference on Research in Information Needs, Seeking and Use in Different Contexts, 13–15 August, 1998, Sheffield, UK.* London: Taylor Graham, 136–149.

Ørom, A. (2000). Information science, historical changes and social aspects: a Nordic outlook. *Journal of Documentation*, 56(1), 12–26.

Patel, V. & Ramoni, M. (1997). Cognitive models of directional inference in expert medical reasoning. In: Feltovich, P., Ford, K. & Hoffman, R., eds. *Expertise in context: human and machine.* Menlo Park, Calif: AAAI Press, 67–99.

Pejtersen, A.M. (1980). Design of a classification scheme for fiction based on an analysis of actual user-librarian communication, and use of the scheme for control of librarians' search strategies. In: Harbo, O. & Kajberg, L., eds. *Theory and application of information research. Proceedings of the 2nd International Research Forum on Information Science.* London: Mansell, 146–159.

Pejtersen, A.M. (1991). *Interfaces based on associative semantics for browsing in information retrieval.* Roskilde : Risø Laboratory.

Pejtersen, A.M. & Fidel, R. (1998). *A framework for work centered evaluation and design: a case study of IR on the web.* Working paper for EU ESPRIT Workshop on Evalution of Multimedia Information Retrieval, Grenoble, 28 March–1 April 1998. (Unpublished)

Reisner, P. (1966). *Evaluation of a 'growing' thesaurus.* Yorktown Heights: IBM Watson Research Center.

Rubinoff, M. (1966). A rapid procedure for launching a microthesaurus. *IEEE Transactions on English Writing & Speech*, 9(1), 8–14.

Saracevic, T., Kantor, P., Chamis, Y.A. & Trivison, D. (1988). A study of information seeking and retrieving I-II-III. *Journal of the American Society for Information Science*, 39(3), 161–216.

Soergel, D. (1974). *Indexing languages and thesauri: construction and maintenance.* Los Angeles, Calif: Melville.

Soergel, D. (1985). *Organizing information.* San Diego, Calif: Academic Press.

Talja, S., Heinisuo, R., Luukkainen, S. & Järvelin, K. (1997). Discourse analysis in the development of a regional information service. In: *Proceedings of the 2nd British-Nordic Conference on Library and Information Studies, Queen Margaret College, Edinburgh, 1997.* London: Taylor Graham, 109–128.

Vakkari, P. (1997). Information seeking in context : a challenging metatheory. In: *Proceedings of an International Conference on Research in Information Needs, Seeking and Use in Different Contexts, 14–16 August, 1996, Tampere, Finland.* London: Taylor Graham, 451–464.

Vakkari, P. (1999). Task complexity, problem structure and information actions. Integrating studies of information seeking and retrieval. *Information Processing & Management*, 35(6), 819–838.

Vakkari, P. (2000). Cognition and changes of search terms and tactics during task performance: a longitude study. In: *Proceedings of the RIAO 2000 Conference.* Paris: CID, 894–907.

Wang, P. (1997). User's information needs at different stages of a research project: a cognitive view. In: Vakkari, P., Savolainen, R. & Dervin, B., eds. *Information seeking in context.* London: Taylor Graham, 307–318.

Wilson, T.D. (1999). Models in information behaviour research. *Journal of Documentation*, 55(3), 249–270.

Winther Jørgensen, M. & Philips, L. (1999). *Diskursanalyse som teori og metode.* Roskilde: Roskilde Universitetsforlag.

Yang, S. (1997). Information seeking as problem solving using a qualitative approach to uncover the novice learner's information-seeking process in a perseus hypertext system. *Library and Information Science Research*, 19(1), 71–92.

(Revised version received 4 June 2001)

Table 9. *Structure and relationships*

Relationship type	*Full name*	*Adm. label*	*Comments*
Note fields	Scope note	SN	Contains a definition of the term and other additional information when appropriate
	Location	URL	Address (web page or Net address) to the external/internal relation or document
	ADM notes	NB	Administrative comments field used by the thesaurus manager
Equivalence relationships	Use/Synonyms & used for	USE/UF	Pointer from terms expressing the same concepts as the preferred term (synonyms or near-synonyms) to the preferred term
	Use full name/Acronym	GO/ACR	Pointer from acronym(s) to the preferred term
	Use Lu number/Generic (INN) name	SEE/GN	Pointer from the generic name of the Lundbeck compound to the Lu number (the preferred term)
	Use the Lu number/Trademark	TO/TM	Pointer from the trademark of the Lundbeck compound to the Lu number (the preferred term)
Hierarchical relationships	Broader terms/narrower terms	BT/NT	Hierarchically superior/inferior terms, clarifying the context and meaning of the concept and suggesting alternative (more general or more specific) terms
	GXP system/GXP Term	GXP/GXT	Broad concepts (e.g. a GXP quality system or topic) describing a certain group of terms related to the particular GXP quality system or topic
Associative relationships	Related terms	RT	Terms associated or related to the concept, clarifying the context and usage of the concept within the company and suggesting alternative concepts to search for
Classification	Subject category	SC	Each term is classified in accordance with a predefined set of subject categories

ELSEVIER

Library Collections, Acquisitions, & Technical Services 24 (2000) 205–215

Library Collections, Acquisitions, & Technical Services

Finding the needle: controlled vocabularies, resource discovery, and Dublin Core

Kuang-Hwei (Janet) Lee–Smeltzer*

Bibliographic Control and Electronic Resource Services, Colorado State University Libraries, Colorado State University Libraries, Fort Collins, CO 80523-1019, USA

Abstract

The phenomenal growth of digital resources on the Internet, their lack of organization, and the deficiency of the search tools currently available, make searching for information on the Internet comparable to looking for the proverbial "needle in a haystack." Developing more effective means of resource discovery and retrieval on the Internet is increasingly necessary. Dublin Core (DC), a newly developed metadata set for resource description, has the potential for providing more effective resource discovery. One major obstacle remains, however: the lack of a systematic approach to subject access. This paper discusses the need for applying controlled vocabularies to enhance the discovery of document-like objects on the Internet and outlines some options for such a process in a distributed environment, with an emphasis on the enhancement of DC with controlled vocabularies. © 2000 Elsevier Science Ltd. All rights reserved.

Keywords: Cataloging; Metadata; Dublin Core; Controlled vocabularies; Resource discovery

1. Introduction

Many metaphors have been used to describe the phenomenal growth of digital resources on the Internet and the lack of effective ways to discover and access them. The Internet has been compared to a mega bookstore where books are piled everywhere: on tables, chairs, floors, with new ones being continuously added to the piles and with their front matter torn off so there is no way to identify the authorship or what the books are about [1]. Another has

* Corresponding author. Tel.: +1-970-491-1849; fax: +1-970-491-4611.
E-mail address: jleesmeltzer@manta.colostate.edu (K.H. Lee–Smeltzer).

1464-9055/00/$ – see front matter © 2000 Elsevier Science Ltd. All rights reserved.
PII: S1464-9055(00)00131-7

described the Internet as "a library in which authors shelve their own books, haphazardly, rewrite them overnight, and move them from place to place without warning" [2]. Most of these metaphors center around the theme of chaos. The chaos is escalating. In April 1998, it was estimated that there were 320 million indexable pages on the Web [3]. By July 1999, the estimated number of Web pages increased to 800 million [4].

The sheer amount of information on the Internet, the lack of organization, and the deficiency of the searching tools currently available—directories (such as Yahoo®) and search engines (such as AltaVista), all contribute to the ineffectiveness of resource discovery on the Internet. The problems and deficiencies of both are well documented in the literature. Although directories, such as Yahoo®, offer some sort of categorization, their subject categories often lack cross-references and the list can get long and unwieldy [5]. It is costly and labor-intensive for humans to create subject categories [6].

Neither search engines nor directories cover the entire Internet. A recent NEC Research Institute study found that "most of the major search engines index less then 10% of the Web. Even by combining all the major search engines, only 42% of the Web has been indexed" [7]. Because directories involve human intervention, it is reasonable to assume that they cover even less of the Internet than the search engines. Yahoo®, boasting the largest directory on the Internet, covers only about 1.2 million classified sites as of April, 1999 [8]. The gap is widening as it becomes more difficult to keep up with the increase in the number of Web pages.

Full-text indexing, non-contextual keyword searching, and little in the way of relevance ranking, result in overly large retrieval sets with low relevancy. Terry–Kuny noted "What is apparent to researchers in distributed indexing is that the current strategy of search engines—that is, to indiscriminately harvest whatever they can find and then do selective indexing on those contents—is an unsustainable architecture for retrieval in a billion document universe" [9]. In addition, search results can be radically different from one search engine to the next.

Why are people both within and outside the information science field so concerned with the discovery and retrieval of Internet resources? It is because the Internet has fundamentally changed the concept of publishing and the way information is disseminated. There are valuable information resources on the Internet that warrant more effective and systematic ways for identifying, describing, and retrieving them. This paper outlines some approaches to applying controlled vocabularies for subject access to enhance resource discovery on the Internet with an emphasis on the enhancement of Dublin Core (DC), a newly developed metadata set for resource description. The discussion is limited to document-like objects (DLO) on the Internet.

2. The metadata movement and the discovery of Internet resources

There are many discussions and debates among information professionals over different approaches for resource discovery on the Internet. One such approach is using metadata for description, and it has gained increasing popularity. Metadata is generally defined as data about data. More specifically, it is a structured set of elements that describe an "information package" for the purpose of identification, discovery, and use of information. An information

Table 1
Dublin Core elements[a]

Content	Intellectual property	Instantiation
Title	Creator	Date
Subject	Publisher	Format
Description	Contributor	Identifier
Type	Rights	Language
Source		
Relation		
Coverage		

[a] Source: Weibel S, Kunze J, Lagoze C, Wolf M. Dublin Core metadata for resource discovery. IETF #2413. The Internet Society, September 1998. (Http://www.ietf.org/rfc/rfc2413.txt).

package is defined as an instance of recorded information, e.g., book, article, Internet document, electronic journal, etc. [10].

Metadata for description is only one type of metadata, albeit one that we are most familiar with in the library world. There are metadata for rating and content selection such as PICS (Platform for Internet Content Selection), for rights management and intellectual property (project Interoperability of Data in E-Commerce Systems (INDECS) is investigating rights management metadata from several different approaches) [11], and for administrative purposes (A-Core) [12]. Descriptive metadata ranges from the very rich, highly technical schema such as MARC formats and TEI headers to the relatively simple 15-element set DC. Metadata can be imbedded in the document itself or exist separately.

3. The development of Dublin Core and its potential as an international standard for resource discovery on the Internet

The DC Metadata Element Set is the result of the March 1995 Metadata Workshop sponsored by the OCLC and the National Center for Supercomputing Applications. Initiated by OCLC, various organizations such as the Research Libraries Group, Coalition for Networked Information, Internet Engineering Task Force, UK Office for Library and Information Networking, and the National Library of Australia have participated in the development of DC. It has gained international attention and support from communities who are concerned with resource discovery on the Internet.

A basic set of DC (DC Simple) consists of 15 descriptive elements that can be categorized into three groups indicating the class or scope of information stored in them (Table 1). All elements are optional and can be repeated. Qualified DC (DCQ) allows each element to be further refined by a limited set of qualifiers. For example, "Creator" can be further defined by "Type," such as PersonalName or CorporateName. "Subject" can be further qualified by "Scheme" such as LCSH (Library of Congress Subject Headings), LCC (Library of Congress Classification), etc. DC records can be generated by document creators, automatically by intelligent software agents, or by trusted third parties such as librarians. DC has the potential of being adopted as an international standard for Internet resource description not only because it is the result of international collaboration and consensus, but also because of its

simplicity, extensibility, and interoperability. The DC element set has already been translated into other languages such as Arabic, Chinese, French, German, Spanish, etc. [13].

DC elements are being reviewed for clarification and formatted to conform with the international standard for expressing the semantics of data elements (ISO 11179). A standards proposal will be submitted to NISO (National Information Standards Organization) in 1999 [14]. DC's simplicity promotes wide and general application, whereas its extensibility allows augmentation by different communities according to their specific needs. The interoperability of DC enables resource description records created using DC to be implemented across disciplines and in different environments.

The simple structure of the set promotes general applicability and can be easily created by Internet information suppliers. However, this also suggests a potential problem: lack of consistency in how the data are entered, thus influencing its effectiveness for resource discovery. Recognizing this, a DC User Guide Working Group is in the process of developing guidelines and promoting "best practices" in a non-technical fashion for both DC Simple and DCQ. The use of controlled vocabularies in the Subject element is also encouraged to enhance the search and retrieval of Internet resources.

4. Why controlled vocabularies are still necessary for taming the net: a literature review

Although keyword and full-text searching can be useful in some cases, the biggest challenge for searching and identifying resources on the Internet remains the lack of effective means of providing subject access. Traditional cataloging and library catalogs have long provided an effective means for accessing the print collection. This success is largely due to applying established cataloging rules and controlled vocabularies to records for each item in a library's collection, thus generating comprehensive and more consistent search results. In his keynote address to the Finding Common Ground Conference, Clifford Lynch stated:

> The Web searching engines—Lycos, Alta Vista, and the like—have been identified by many as the future of information organization . . . these systems lack high-quality retrieval capabilities for many applications (when contrasted to systems designed around retrieval from intellectual cataloging, abstracting and indexing by human beings, such as online library cataloging or good A&I database implementations) [15].

For the purpose of this paper, controlled vocabularies are defined as classification schema or lists of terms and phrases selected to express certain concepts that are applied to records in a database in a consistent manner. A classification scheme places subjects in categories and assigns a set of numbers, letters, symbols, or a combination of all of the above, in a hierarchy. Controlled vocabulary can be universal schema such as DDC (Dewey Decimal Classification), LCC, LCSH, schema specific to a discipline, such as the National Library of Medicine classification and subject headings (MeSh), or locally devised systems.

A literature review indicates that controlled vocabulary is crucial for effective search and retrieval of Internet resources [16,17,18,19]. It is also interesting to note that according to the April 1998 Search Engine Report, directory-style architecture for portal sites is gaining in

popularity with commercial search engines. For example, AltaVista and other search engines have added directories to their services. "[D]irectories are on the rise mainly as a response to users continuing to perform overly broad searches, which can leave them feeling lost among a sea of off-target results" [20]. Even with the problems associated with directories mentioned previously, searching through rough subject categories is still considered a more structured approach than use of full-text keyword searches. Furthermore, the larger the files being searched and the broader the subject coverage, the greater the need for controlled vocabularies [21]. In a language rich in synonyms, such as English, without the controlled form for access to a particular subject term, users will have to search for every variation of that term or run the risk of retrieving too many entries and missing important ones, not even knowing what they have missed [22].

Lassila pointed out that "[m]achines (in this case, search engines) cannot understand a natural-language document and thus cannot always extract specific information from the document such as author, publication date, or topic." [23] "[K]eywords thus perform poorly in situations where a search index covers multiple subject areas, as is the case with the Web." [24] Furthermore, Lassila suggested that it would be useful if other means of searching were available to us in addition to full-text search, such as the author of the document, the date it was published, and the specific topic it discussed, although the topic might not be expressed as a particular word in the document itself [25]. This suggestion implies some form of content analysis and application of controlled vocabulary to these documents. "The difference between keyword searching and controlled vocabulary searching is the distinction between mechanistic access to isolated terms and intellectual access to unified concepts. This is the key to the utilization of the Internet . . . the presence of a controlled vocabulary greatly enhances the usability of any information retrieval system" [26].

5. Scenarios for applying controlled vocabularies for subject access to Internet resources in a shared environment

Analyzing the subject content of a work and assigning controlled vocabulary are the most labor-intensive and costly elements of the cataloging process. The dynamic nature of resources on the Internet adds further difficulty to the already labor-intensive process. Options for better subject access, however, need to be developed and examined for more effective resource discovery. The following approaches outline some possible scenarios for accomplishing this. They are by no means exhaustive. Some of these approaches are already being used.

5.1. The traditional cataloging approach

Libraries have not been collecting all the print materials that have been published, only the selected ones that fit the mission and the purposes of the institution. The same should apply to the electronic resources on the Internet. As Erik Jul has pointed out, not all Internet resources are worth collecting and cataloging. However, if libraries of different types and sizes with diverse subject emphases and interests follow the same systematic approach as

they have for collecting and making access to print materials, the "most worthy" Internet resources will be identified, described, and accessible for the users. Even so, this approach does not address a bigger problem: a very small percentage of Internet resources will get this labor-intensive treatment.

Libraries will take a more proactive role in selecting "worthy" Web resources for inclusion in their collection in accordance with their collection development policies. Cataloging records with controlled vocabularies for subject access can be created for these resources using the current cataloging rules and MARC formats. These records can be input into the bibliographic utility for sharing with other libraries and for inclusion in the local online catalog. The process is easily integrated into the cataloging workflow. In addition, the OCLC Intercat project and the cataloging community's continuous efforts to revise the cataloging rules and MARC formats to accommodate some of the fundamental differences posed by electronic resources on the Internet will help shape the cataloging standards and practices that are still evolving.

5.2. The DC metadata approach

Although the MARC record provides a rich format for resource description, its use is also labor-intensive and requires special technical knowledge of the rules for cataloging and MARC encoding. DC was developed specifically for describing document-like objects on the Web. Once DC becomes a standard and is widely implemented, it may prove to be a more suitable scheme for resource discovery on the Internet. The creation of a simple DC record does not require special training and can even be done automatically with a template and author-supplied information. DC can be encoded in HTML or XML, both of which are, or will be, more widely used than MARC formats that are usually limited to the library community. For resources on the Internet, there are other considerations such as intellectual property, rights management, and administrative information that may be handled more effectively by DC (or other metadata formats) than by MARC. For these and other reasons, DC may be more attractive than MARC in the commercial marketplace for developing electronic resources.

The initial intention for DC is to have a simple DC record created by the information provider at the point of posting the resource on the Internet. Trusted third parties such as librarians can then enhance the record. Using the scheme qualifier under the element "subject," the record can be enhanced by various controlled vocabularies such as DDC, LCC, LCSH, etc. To share the enhanced record in a distributed environment, a central registry or database becomes necessary. In fall 1998, OCLC announced the experimental Cooperative Online Resource Catalog (CORC) project to explore the cooperative creation and use of metadata primarily for online resources. Based on the WorldCat model for print materials, CORC will explore the feasibility of having different record structures, namely MARC and DC, in the same database. By using Resource Description Framework (RDF)-compliant Extensible Markup Language (XML), HTML, and MARC, records can be imported into or exported from CORC. RDF is a standard recommended by the World Wide Web Consortia (W3C). It provides a framework for exchanging metadata of many varieties. The CORC system can harvest online resources and create

simple descriptions of Web resources automatically. Participants of CORC can create and contribute records to the database in the same fashion as OCLC member libraries share records in WorldCat. Records can then be exported for local use. CORC has plans to support authority control and automatic assignment of classification numbers and subject headings. CORC is another step forward in the effort of improving discovery and access to resources on the Internet. If successful, it also has significant implications for the library world. The current OPAC technology is not capable of containing and processing different types of data structures. The CORC system can be modeled for the development of future generations of library OPACs.

There are many other DC projects in North America, Europe, Australia, and Nordic countries: University of California—Berkeley's Digital Library Catalog (http://sunsite.berkeley.edu/Catalog), the Gateway to Educational Materials (GEM) initiated by the US Department of Education and the National Library of Education (http://gem.syr.edu), the Nordic Metadata I project funded by NORDINFO from October 1996 to June 1998 (http://linnea.helsinki.fi/meta/), and the PANDORA Project at the National Library of Australia for building an online archive (http://www.nla.gov.au/policy/pandje97.html). The hope for a wide implementation of DC has yet to become a reality. Deployment of DC has not reached the critical mass for most commercial search engine developers to be interested in incorporating DC into their search algorithms. Early experiments with and implementations of DC have been mostly carried out by libraries and research institutions that collect or create scholarly information as illustrated by the examples named above.

5.3. Automatic subject analysis and assignment of controlled vocabularies

Classification and application of subject headings are the most labor-intensive and costly parts of document description. They are also the processes most difficult for humans to apply in a large-scale distributed information environment such as the Internet. Can options be developed to automate the subject analysis and the assignment of controlled vocabularies?

No computer programming at present can fully replace the human intellectual process of assigning controlled vocabularies. None of the past experiments with automatically assigning classification numbers and subject headings was very successful. With the advancement of computing technology and the development of artificial intelligence and natural language processing, however, it is conceivable that in the future such programs could be designed to render acceptable results.

It is possible for a sophisticated computer program to examine the entire contents of an electronic document, or parts thereof, such as the abstract, summary, and the bibliographical references, to determine words that are significant within context according to some algorithm. The missing piece in this scenario is a program that is capable of matching the result to an existing list of controlled terms or classification scheme and assigning the "appropriate" subject headings and/or classification number automatically, and with a high degree of accuracy.

It is important to note that several research and experimental projects in this area are currently underway. These projects approach the automating process with different models (computer linguistic methods, neural network models, etc.), and some are showing promising

results. OCLC Scorpion is a research project designed to build "tools for automatic subject recognition by combining library science and information retrieval techniques" [27]. An electronic document is input as a query against the Scorpion Dewey Databases, using the SMART (System for Manipulating And Retrieving Text) ranked retrieval algorithm. The program returns a ranked list of potential Dewey numbers as subjects. The list can then be reviewed by a human cataloger. Scorpion is used for generating classification numbers and subject headings for OCLC NetFirst, a database of Web resources available via the OCLC FirstSearch and OCLC Cataloging services.

Another interesting project is the development of "Entry Vocabulary Modules" (EVMs) at the University of California–Berkeley, School of Information Management and Systems. The project experimented with mapping natural language queries to controlled vocabularies such as the LCSH via an association dictionary. Although the experiment was conducted in specific databases as well as the catalog of the University of California's online library system MELVYL [28], further research could be conducted to test the feasibility of applying this experiment's concept and methodology to resources on the Internet.

5.4. Subject access to resources on the Internet as one virtual library via subject gateways

Subject gateways are sometimes referred to as subject-based information gateways, subject trees, etc. Many such gateways have been developed to enhance resource discovery and to provide organization of resources on the Internet in specific subject areas. Some of them arrange resources by classification schema and others by subject headings. Most of these subject gateways link to resources that have been through a human reviewing or selection process. Examples of subject gateways include: CyberStacks(sm) (http://www.public.iastate.edu/~CYBERSTACKS/); Art, Design, Architecture & Media Information Gateway (ADAM) (http://www.adam.ac.uk/adam/index.html); and Organizing Medical Networked Information (OMNI) (http://omni.ac.uk/).

Although similar to traditional cataloging—the assignment of classification and subject headings is still performed by humans—subject gateways take a different approach from the traditional library OPAC in that they treat the entire Internet as one virtual library and, in many cases, provide additional searching facilities and subject-related services [29]. Some gateways focus on one single subject area whereas others cover several different disciplines. These gateways provide better subject access to quality resources on the Internet. However, as more and more subject gateways are created and multiple gateways are devoted to the same subject, searching all of them is time-consuming but selecting which ones to search can be equally difficult if there is no cross searching. This is particularly problematic when searching for information in interdisciplinary areas. In an effort to improve the situation and develop a system whereby cross-search in multiple gateways can be executed by a single query, Kirriemuir et al. described a query routing and forward knowledge approach using the Common Indexing Protocol to facilitate cross-searching in multiple gateways [30].

5.5. Create online thesauri for Internet resources

The association dictionary in Berkeley's EVMs project can be viewed as an online thesaurus. In the first stage of the experiment, an algorithm was developed by studying and learning the association between the lexical clues found in the cataloging records and the subject headings assigned by human catalogers [31]. The advantages of an online thesaurus are that it can be modified and updated more easily and automatically. The results of the mapping can either be transparent to the searcher or a ranked list of choices of broader, narrower, or related terms for the searcher to further refine the search. The thesaurus may have the capability of mapping the natural language search terms to more than one controlled vocabulary scheme as well as linking among different controlled vocabulary schema. The creation and maintenance of such an online thesaurus for Internet resources, however, can be a tremendous task. Although most of the process may be done automatically, human intervention will still be needed to decide the structure and mapping mechanism of the thesaurus. If different online thesauri are created for different disciplines, is it beneficial to provide links among them? How can the appropriate links be created and maintained?

6. Conclusion

We have reviewed various efforts from many different communities (library science, information science, computer science) that attempt to better organize Internet resources and provide more effective methods for resource discovery. These approaches range from traditional cataloging in the library world to the recent metadata movement and the development of the Resource Description Framework for building an architecture that enables the co-existence of different data structures. However, it is clear that one size does not fit all. Traditional library cataloging, MARC records, and the various metadata schema all have their places in the sense- and order-making business of the Internet. Some resources may warrant the most labor-intensive and detailed description, whereas for others a description based on a simple metadata scheme such as DC Simple will suffice.

Librarians, catalogers specifically, have long developed the expertise in organizing information stored in physical media and making them accessible in a coherent and consistent manner through the library catalog. Coherence and consistency are achieved by applying established cataloging rules and the use of controlled vocabularies. Although keywords and full-text indexing can be useful in some situations, controlled vocabularies are still necessary for effective resource discovery in large distributed databases with broad subject coverage such as the Internet. The skills of analyzing the intellectual content of a work, classifying it, applying subject headings to it, and arranging the library's collection logically are easily transferred to the organization of information stored electronically. There are issues, however, yet to be resolved and standards to be developed for organizing and managing this new medium.

In addition to using our expertise and sharing our experience in knowledge and information management, librarians need to be actively involved in the development of many other metadata schema, as well as innovative ways of making resource discovery on the Internet

more effective. Only through collaborative efforts from different communities can order be achieved on the Internet. Clifford Lynch noted ". . . the librarian's classification and selection skills must be complemented by the computer scientist's ability to automate the task of indexing and storing information. Only a synthesis of the differing perspectives brought by both professions will allow this new medium to remain viable" [32]. In this way we can reduce the difficulty in finding and retrieving the needles in the Internet haystack.

References

[1] Taylor AG. Perspectives on the subject of subjects. Journal of Academic Librarianship 1995;21:485.
[2] Mandel CA, Wolven R. Intellectual access to digital documents: joining proven principles with new technologies. In: Ling-yuh WP, Cox BJ, editors. Electronic resources: selection and bibliographic control. New York: Haworth, 1996. p. 26.
[3] Search engine sizes. Search Engine Watch, 1999. http. //searchenginewatch.com/reports/sizes.html.
[4] Search engine sizes. Search Engine Watch, 1999. http. //searchenginewatch.com/reports/sizes.html.
[5] Vellucci SL. Options for organizing electronic resources: the coexistence of metadata. Bulletin of the American Society for Information Science 1997;24:16.
[6] Filman RE, Sangam P. Searching the Internet. IEEE Internet Computing 1998;2:21.
[7] Dunn A. Most of web beyond scope of search sites. Los Angeles Times Business, 1999. Http://www.latimes.com/HOME/BUSINESS/UPDATE/lat search990708.htm
[8] Directory Sizes. Search Engine Watch http://www.searchenginewatch.com/reports/directories.html.
[9] A Note on Internet Search Engines and Metadata (Draft December 11, 1998) posted on the Knowledge Access Management Discussion Forum KAM @OCLC.org on 1/15/99.
[10] Taylor AG. The organization of information. Englewood, CO: Libraries Unlimited, 1999. p. 246.
[11] Rust G. Metadata: The right approach: an integrated model for descriptive and rights metadata in e-commerce. D-Lib Magazine, 1998. http://www.dlib.org/dlib/july98/rust/07rust.html.
[12] Internet Draft: http://metadata.net/admin/draft-iannella-admin-01.txt.
[13] Weibel S. The state of the Dublin Core metadata initiative April 1999. D-Lib Magazine 1999;5. http://www.dlib.org/dlib/april99/04weibel.html.
[14] Ibid.
[15] Lynch CA. Keynote address–finding common ground. In: LaGuardia C, Mitchell BA, editors. Finding common ground: creating the library of the future without diminishing the library of the past. New York: Neal–Schuman Publishers, 1998. p. 7.
[16] Desai BC. Supporting discovery in virtual libraries. Journal of the American Society for Information Science 1997;48:190–204.
[17] Micco M. Subject authority control in the world of Internet: Part 1. LIBRES: Library and Information Science Research 1996;6. http://www.bubl.ac.uk/journals/lis/kn/libres/v06n0396/micco 1.html.
[18] Micco M. Subject authority control in the world of Internet: Part 2. LIBRES: Library and Information Science Research 1996;6. http://www.bubl.ac.uk/journals/lis/kn/libres/v06n0396/micco 2.html.
[19] Taylor AG. Perspectives on the subject of subjects. Journal of Academic Librarianship 1995;21:484–91.
[20] Directories Take Center Stage, The Search Engine Report, 1998. http://searchenginewatch.internet.com/sereport/9805-directory.html.
[21] Hagler R. The bibliographic record and information technology, 3rd edition. Chicago: American Library Association, 1997. p. 263.
[22] Tillett BB. International shared resource records for controlled access. ALCTS Newsletter Online: From Cataloging to Gateway 1998;10. http://www.ala.org/accts/alcts news/v10n1/gateway.html).
[23] Lassila O. Web metadata: a matter of semantics. IEEE Internet Computing 1998;2:30.
[24] Ibid.
[25] Ibid.

[26] Connaway LS, Wallace DP. Organized access to engineering internet resources using indexing principles. In: LaGuardia C, Mitchell BA, editors. Finding common ground: creating the library of the future without diminishing the library of the past. New York: Neal–Schuman Publishers, 1998. p. 393.
[27] Shafer K. Scorpion helps cataloging the web. June 99. http://orc.rsch.oclc.org:6109/b-asis.html.
[28] Norgard B. Entry vocabulary models and agents. June 28, 1998. Rev. July 14, 1998. http://www.sims.berkeley.edu/research/metadata/agents.html. Project homepage: http://www.sims.berkeley.edu/research/metadata/).
[29] Kirriemuir J, Brickley D, Welsh S, Knight J, Hamilton M. Cross-searching subject gateways: the query routing and forward knowledge approach. D-Lib Magazine Jan. 1998. http://www.dlib.org/dlib/january98/01kirriemuir.html.
[30] Ibid.
[31] Plaunt C, Norgard BA An association-based method for automatic indexing with a controlled vocabulary. Journal of the American Society for Information Science 1998;49:888.
[32] Lynch C. Searching the Internet: combining the skills of the librarian and the computer scientist may help organize the anarchy of the internet. Scientific American 1997;276:52.

Argus Center for Information Architecture

Extracting Value From Automated Classification Tools

The Role of Manual Involvement and Controlled Vocabularies

By Kat Hagedorn, Argus Associates

March 2001

http://argus-acia.com/

Copyright 2001, Argus Associates, Inc. All Rights Reserved.

Table of Contents

Introduction 3

Automated Classification 101 4

- Definition of Classification 4
- Benefits of Classification 4
- Automated Classification Tools 5

Adding Humans into the Mix 6

- Range of Human Involvement 6
- Benefits and Limitations of Manual and Automated Approaches 7

Using Controlled Vocabularies 9

- Definitions of Controlled Vocabulary and Thesaurus 9
- Benefits of Controlled Vocabularies and Thesauri 10
- Incorporating Controlled Vocabularies 11
- Obtaining Controlled Vocabularies 11
- Issues to be Aware of During Development 12

A Call for Testing 13

- Testbed Features 14
- Benefits of Testing 14

References 15

About the Author 16

About the Argus Center for Information Architecture 17

Copyright 2001, Argus Associates, Inc. All Rights Reserved.

INTRODUCTION

Automated classification has been trumpeted by vendors as the solution to the problems we face in managing our huge and growing information spaces. Most of us feel like clutching this solution as a drowning person would a life vest because it offers a way of re-gaining control over this information.

We do need a solution, but in fact, as in most things, this type of solution works only if you have a plan and a process for using it. A plan and a process are developed from understanding the environment of your information space—in other words, your business context, content and users. Examples of questions to ask about implementing an automated classification solution in your environment include:

- **Context.** Is there buy-in at the top levels of management for automated classification and for providing the appropriate resources for it? Do you have staff members who fully understand the nature of your business and can translate that into an appropriate classification?
- **Users.** Are your users primarily searchers or browsers? Are they interested in categories that are topical, task-oriented, geographical or audience-based, or do they need customized organizational methods?
- **Content.** Does your information space contain content that is all one type (i.e., homogenous) or different types? Is it in multiple languages?

As part of developing this plan and process, you will also need to:

- **Choose the mix of manual and automated classification.** Do you need humans to create rules used to classify the information? Will you expect the tool to suggest classifications for humans to review? What level of quality control will be necessary?
- **Choosing the controlled vocabularies you use with the automated classification tools.** Will you need an author vocabulary as well as a subject vocabulary? Do you have the resources to build the controlled vocabulary from scratch? Will a thesaurus be more suitable?

This white paper will help answer these questions and others. It discusses the types of automated classification tools that are available, how manual classification fits into the mix, why you need a controlled vocabulary to do the job right, and asks how we can move towards community-wide testing of these automated classification tools.

COPYRIGHT 2001, ARGUS ASSOCIATES, INC. ALL RIGHTS RESERVED.

AUTOMATED CLASSIFICATION 101

Automated classification has been described in a multitude of ways, depending on the point of view of the author (e.g., academician, vendor). A recent excellent discussion is provided in Katherine C. Adams' article "Word Wranglers" (1), which includes a description of the tools available, their features and the techniques used for clustering information.

The following takes a step back and explains what classification is in general, why it is useful and then briefly summarizes the automated classification tool techniques. It is not our intention to discuss details of any of these tools—Adams' article goes into depth on this topic.

DEFINITION OF CLASSIFICATION

Classification is the process by which information, whether in document or data format, is clustered together to make it easier for the user to find it. (For the scope of this white paper, the terms "classification" and "categorization" can be used synonymously.)

For instance, while classifying a collection of documents about food, you might want to cluster documents by subject—e.g., those that discuss "methods of cooking," those that discuss "kitchen tools"—depending on how useful your audience would find that type of classification.

BENEFITS OF CLASSIFICATION

In a nutshell, classification assists people who are:

- **Browsing**. For example, a user may browse through a classified structure (e.g., Men's Clothing: Shirts: T-Shirts) to find relevant information—this structure reflects clusters created by classification.
- **Searching**. For example, a user searching for "cars" retrieves all information that is clustered around this topic. The search engine retrieves these clusters and returns these results to the user.
- **Managing**. For example, an editor looking for the appropriate place to put a piece of information can use a classified structure to help suggest places. The resulting placement of information forms clusters.

COPYRIGHT 2001, ARGUS ASSOCIATES, INC. ALL RIGHTS RESERVED.

The end result of classification is structures that are commonly known as ontologies, taxonomies, hierarchies, controlled vocabularies or thesauri—depending on the discipline involved. A superb article by Dagobert Soergel (2) reflects on the use of these names and notes that one of the main features of any classification structure is to support information retrieval. In this white paper, I will focus on the use of specific types of these structures—i.e., controlled vocabularies and thesauri—to support more efficient and relevant information retrieval.

AUTOMATED CLASSIFICATION TOOLS

Automated classification is the process by which technology is used to create clusters. Some of the most popular vendors that have marketed stand-alone automated classification tools or add-ons to their search or content management tools are Autonomy, Inktomi (formerly Ultraseek), Interwoven (formerly Metacode), Mohomine, Semio and Verity.

These tools cluster information by using one of the following techniques:

- **Statistical clustering**. Employs algorithms to cluster information. Popular methods include term co-occurrence analysis and neural networks. This technique is dependent on the information in the collection—it classifies using the terms in the collection exclusively. Vendors using this technique are Autonomy, Interwoven, Mohomine and Semio.
- **Rules-based clustering**. Necessitates the creation of IF-THEN statements that define the clusters of information. This technique builds a classification structure that can be reviewed and edited manually, while populated automatically. It isn't dependent on the information in the collection—a new collection could use the same statements. Vendors using this technique are Inktomi and Verity.

An additional technique, designed for improving the above techniques, is:

- **Training**. Compares to-be-classified information with previously well-classified information, and collects these together in clusters. Training can happen at the beginning of the classification process or iteratively throughout. Vendors using this technique are Autonomy, Inktomi and Mohomine.

These automated classification tools work with varying degrees of success. This success depends on the involvement of humans and the incorporation of controlled vocabularies, each discussed in the next two sections.

COPYRIGHT 2001, ARGUS ASSOCIATES, INC. ALL RIGHTS RESERVED.

Adding Humans into the Mix

A number of automated classification tool vendors, pundits and those working in this field have noted that fully automated classification of information is not the complete answer. (1,3,4,5) They state that a semi-automated solution is more appropriate. Peter Morville puts it most succinctly in his ACIA "Little Blue Folders" article: "The key to success in designing information architecture solutions for really large web sites and intranets is to intelligently combine manual AND automated approaches." (6)

Recently, there have been detractors concerned that the cost-benefit gap may be too large, suggesting that involving humans is an unnecessary time expenditure when machines are capable of doing the entire process. (7) I would respectfully disagree. Although there haven't been formal evaluations of this, in my experience manual intervention is necessary, especially for the conceptual parts of the process.

Good indexers and catalogers know the difficulties of developing appropriate clusters of information. Often it isn't enough to just pull out terms that reside in documents; concepts that aren't stated verbatim within a document also need to be reflected in the classification. Computers are not able to understand the meaning in documents or the context surrounding them. As stated by Nancy Mulvany, "computers are capable of automatically manipulating the text of a book in a variety of ways. Yet the computer is incapable of exercising the type of judgment and interpretation applied by experienced indexers." (8)

Range of Human Involvement

The use of humans in the classification process runs the gamut from minimal involvement to full-fledged participation.

Humans can be involved with automated classification tools by:

- **Creating the classification.** Developing all or part of the structure to be used during the automated classification process. For instance, manually creating the top-level clusters of the classification structure and having the tool automatically create the bottom-level clusters. This structure can be newly created or tweaked from out-of-the-box versions, and should be predicated on a controlled vocabulary or thesaurus (see next section).

Copyright 2001, Argus Associates, Inc. All Rights Reserved.

- **Developing the classification rules.** Manually creating and editing the rules that govern the clustering of information. These rules should also be predicated on a controlled vocabulary or thesaurus.
- **Training the collection process.** Moving information from one cluster to another, when necessary, so that the tool learns what information should exist in specific clusters.
- **Implementing suggestions.** Reviewing options provided by the tool for classifying information, analyzing these and using them or choosing other options to best cluster the information.

Throughout the automated classification process, humans should be:

- **Performing quality control.** Reviewing clusters and the information in the clusters after the tool has done its work, and making any changes. Hopefully, the tool also allows humans to perform some of the above listed functions (e.g., creating the classification, training the tool), in order to reduce the need for extensive manual quality control.
- **Testing the classification.** Designing methods for testing the usability (e.g., appropriateness, ease of use) of the resulting classification structure, running tests and analyzing the results. Analysis should provide you with recommendations for improvement to the classification structure and possibly improvements to the classification process.

Benefits and Limitations of Manual and Automated Approaches

To fully understand the benefits and limitations of choosing manual and automated classification methods, you will need to develop a table similar to the following:

Approach	Benefits	Limitations
Manual	▪ Better understanding of conceptual nuances ▪ Understanding of company environment ▪ High quality ▪ High consistency ▪ Ability to modify structure ▪ Ability to classify images and applications	▪ Slower ▪ Perceived as resource-hungry ▪ Perceived as high cost

Copyright 2001, Argus Associates, Inc. All Rights Reserved.

APPROACH	BENEFITS	LIMITATIONS
Automated	▪ Perceived cost savings ▪ Perceived as fewer resources required ▪ Frees humans from routine tasks ▪ Minimizes time to complete classification process	▪ Uncertain quality ▪ Uncertain consistency ▪ Overhead for technology and its expertise ▪ Inability to deal with exceptions to rules ▪ Often needs fairly homogeneous collections ▪ Often needs large collections ▪ Often needs full-text collections (not records)

Your environment aside, the most pressing issue of your classification efforts is the quality and consistency of the classification process. The following section addresses how implementing controlled vocabularies assists in limiting poor quality and inconsistency in your classification.

COPYRIGHT 2001, ARGUS ASSOCIATES, INC. ALL RIGHTS RESERVED.

Using Controlled Vocabularies

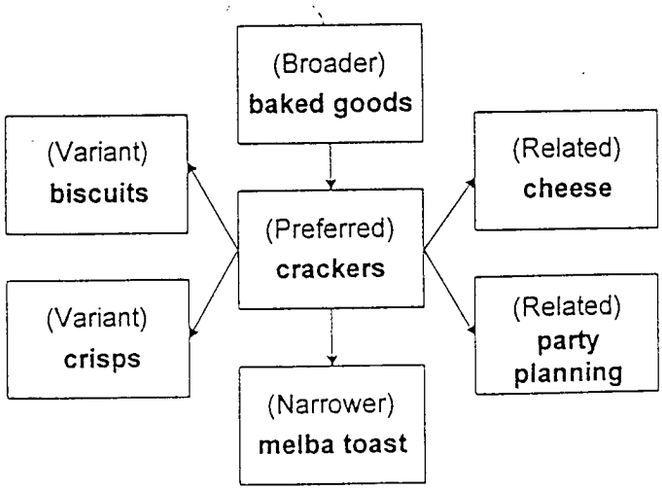
Here's where the users matter the most. You've developed a system that adequately classifies the information you have in your information space, based on the resources and capabilities of your company. You've determined the mix of automated and manual classification and have chosen an appropriate tool. The next step is to identify what the users want.

Most users are interested in directly relevant information (as opposed to all potentially interesting information). Each of the automated classification techniques noted above claims to give users the most relevant information. But if you really want to ensure that users are retrieving what they need, you should be using a controlled vocabulary or thesaurus. Reasons for using these types of classifications are explored in this section.

Definitions of Controlled Vocabulary and Thesaurus

A controlled vocabulary (CV) is a list of descriptive terms that indicates which terms are preferred and which are variants of the preferred terms. A thesaurus is a special kind of controlled vocabulary that also indicates which terms are broader, narrower and related to the preferred terms. For instance, a controlled vocabulary on food might indicate that "crackers" is the preferred term and "crisps" is a variant term. A thesaurus might also indicate that "cheese" is a related term and that "melba toast" is a narrower term. The following diagram illustrates these relationships.

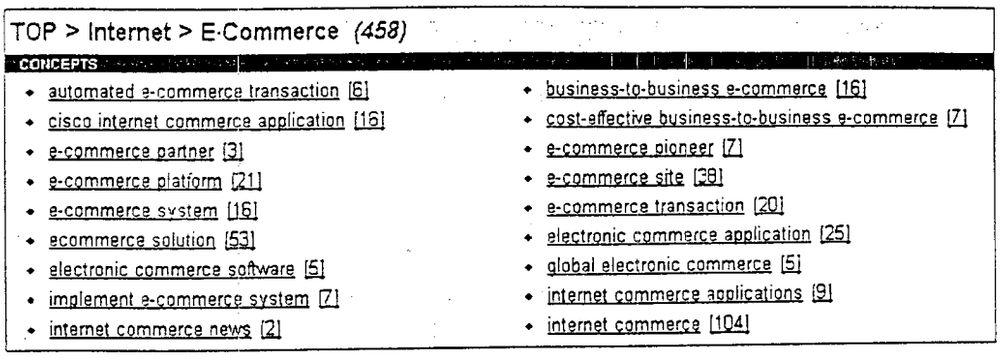
Figure 1. Thesaurus: relationships between terms.



Copyright 2001, Argus Associates, Inc. All Rights Reserved.

Without a CV in place, automated classification tools can produce something like the following:

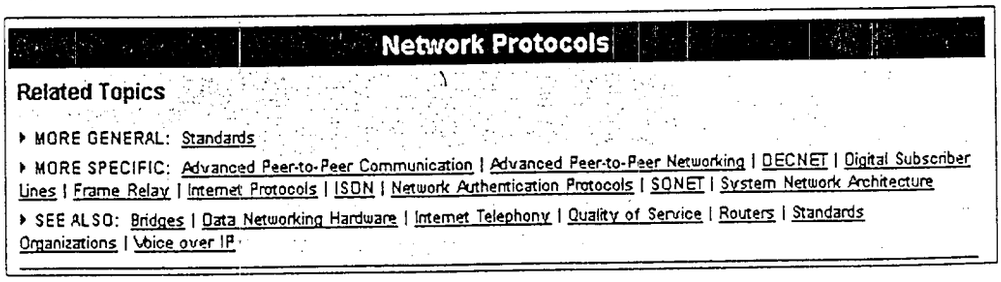
Figure 2. Semio: classification of "e-commerce."



In this example, if a CV is in place, a preferred term for "application" or "software" can be chosen, with the other one listed as a variant term. Then, at least three clusters might be merged into one (e.g., "electronic commerce application," "internet commerce application," "electronic commerce software"). And, using a thesaurus, narrower terms might be used for sub-clustering (e.g., "shopping cart tools," "product finders").

Consequently, information is re-distributed into appropriate groupings based on the CV or thesaurus. The end result is more relevant clusters of information for users—they can find what they need gathered in one place. Using this method, well-classified information can look like the following:

Figure 3. Bitpipe: thesaurus term "network protocols."



In this example, broader, narrower and related terms are provided for "network protocols." This gives the user context for this term and avenues for further exploration—it enhances their information retrieval experience and increases the likelihood that they will find relevant results.

COPYRIGHT 2001, ARGUS ASSOCIATES, INC. ALL RIGHTS RESERVED.

For more examples and further resources about CVs and thesauri, see the online materials from the 2001 ACIA "Synonyms and Taxonomies" seminar (argus-acia.com/seminars).

Incorporating Controlled Vocabularies

So, how do you use CVs with automated classification tools? In essence, anywhere a classification is used, a CV can be used. Options include:

- **Creating** the top levels of the classification with a CV and letting the tool automatically create the bottom levels.
- **Informing** the creation and editing of IF-THEN statements with a CV.
- **Training** by selecting preferred and variant terms from a CV and populating clusters with information that includes those terms.
- **Comparing** suggestions for clustering with a CV.
- **Reviewing** statistically derived clusters, matching them against a CV and changing the clusters appropriately.

Obtaining Controlled Vocabularies

Procuring pre-built CVs can be difficult for a variety of reasons. They can be:

- **Inexistent.** Is there a CV that has been created in your subject domain and that fits how your users look for information?
- **Incomplete.** Does the CV cover all the topics you're trying to cover?
- **Inconsistent.** Does the CV consistently organize items at appropriate levels of detail?
- **Incomprehensible.** Is the CV too technical for your needs?

The bottom line is: How much will the CV need to be tweaked? Sometimes the best approach is to build one from scratch and refine it during the classification process. This means involving more resources (time, money, people), but the end result will be more satisfactory than a pre-built CV.

Another option is to use the text to generate the controlled vocabulary (as opposed to statistically generating clusters that aren't controlled). There has been research done on the use of tools that automatically extract terms from information and then build a thesaurus from those terms. (9) This method is best used in conjunction with manual involvement to ensure the relevancy of the CV. The benefit of this method is that the CV is directly connected to the information being clustered.

COPYRIGHT 2001, ARGUS ASSOCIATES, INC. ALL RIGHTS RESERVED.

See Lou Rosenfeld's recent article "Looking for Metadata in All the Wrong Places" for an accessible discussion of CVs and how to obtain them. (10)

Issues to be Aware of During Development

Developing CVs involves issues that may not be obvious initially. Some of these are:

- **Creating cross-references or double-posting.** It's often useful to provide "see" references (e.g., "cars *see* automobiles") and "see also" references (e.g., "cars *see also* travel") among clusters, or to selectively place information in more than one cluster (i.e., double-post it). This makes the relationships among terms obvious and makes the resulting classification more usable.
- **Building more than one CV.** To reach your primary users and to reflect all your content, you may need to develop multiple types of CVs. Examples include those organized by user role (e.g., engineer, administrator), topic (e.g., crackers, cookies), action or task (e.g., selling, buying) and geography (e.g., France, England).
- **Building CVs in more than one language.** If your collection contains considerable amounts of information in multiple languages, you should have a CV for each language. This often involves more than simple translation of terms, since concepts, jargon and relevancy of terms differ among languages.
- **Building a thesaurus.** If you have the resources, a thesaurus can greatly help your users and your business. You will be offering users navigational assistance (e.g., broader, narrower, related terms), which can help them generate more relevant results. Managing a classification structure that includes more relationships among terms can be a very powerful avenue for translating your business strategy—e.g., promoting products using broader clusters ("you're interested in 'construction tools,' you may want to see everything related to 'home improvement'").

Copyright 2001, Argus Associates, Inc. All Rights Reserved.

A Call for Testing

I've discussed the need for manual involvement in your classification process and I've discussed the need for controlled vocabularies in that process. What I think is missing in these discussions is testing of these variable, and sharing of the results with the wider community.

Situations exist in which testing can occur, but inevitably there are limitations to doing this kind of testing. These situations include:

- **Partnering with an automated classification tool vendor** so you can run pilot tests with actual data using their tool. The limitation is the partnership agreement and how that affects your company's strategy.
- **Obtaining a trial version of a tool** to run against actual or contrived data. The limitation is that the trial version may limit you to a certain number of documents, not allow the integration of controlled vocabularies or not allow you to customize the classification process.
- **Contracting with a vendor to do a full test of their tool**, which usually involves the assistance of consulting engineers from the vendor. The limitation is the added expense of paying for the consultants just to see whether you want to use the tool.

I feel that there is a greater opportunity out there—an opportunity to provide the community with real data that has been tested in a variety of ways using a variety of automated classification tools, with results that can be shared and compared without sharing proprietary vendor information with the community. In other words, I feel there is an opportunity to form an international collaboratory testbed for these types of tools.

The Text Retrieval Conferences (TREC) project provides an excellent example of a collaboratory testbed (trec.nist.gov). Each year, the project provides data sets that interested parties can use to run against new technology to see where that technology succeeds and fails. It's not a giant leap to assume that something like this could be done to test automated classification tools. Since TREC's environment is set up for new technology and not established technology, it may not be the best venue for the proposed testing.

COPYRIGHT 2001, ARGUS ASSOCIATES, INC. ALL RIGHTS RESERVED.

TESTBED FEATURES

In order for use of the testbed to be most effective, the community needs to enforce:

- **Description of the information** in the testbed. What type of information are you testing (e.g., engineering, religious studies)? Is it mostly text or records-based? Is it highly conceptual content or fact-based data?
- **Information in multiple languages.** This will assist in reaching the largest community of testers.
- **Selection of controlled vocabularies** to run against information. Multiple types of CVs are a necessity. Multiple language CVs will be necessary for information in multiple languages.
- **Ability to perform the test in a multitude of ways.** For instance, running text-based English language content with multiple topic domain CVs using three selected tools, or running records-based French and English language data without a CV through five selected tools.

BENEFITS OF TESTING

The testbed would benefit both the automated classification vendors and those interested in using the tools. It can assist in:

- **Selling the need for the tool to the leadership.** Quantifiable effects are easier for upper management to buy into.
- **Providing a way for content managers to compare tools** and see what types of tools work best for their types of data and resources.
- **Comparing manual and automated methods of classification.** The level of human involvement could be tested under a variety of conditions.
- **Getting feedback from users.** Testing with users can provide subjective and objective data on which tools work in which environments.
- **Selling the tool to you.** Vendors with quantifiable data on their tool's performance can use this to market the tool.

As a consultant, I also find the opportunity to share results from prior and present engagements to be a strong selling point for future clients. There is a dearth of tangible outcomes in the field of information architecture. It would be useful to have concrete results showing that controlled vocabularies are effective using certain tools and certain types of information—this would benefit the client and the entire information architecture community.

COPYRIGHT 2001, ARGUS ASSOCIATES, INC. ALL RIGHTS RESERVED.

REFERENCES

1. Adams, K.C. "Word Wranglers: Automatic Classification Tools Transform Enterprise Documents from 'Bags of Words' into Knowledge Resources." (www.intelligentkm.com/feature/010101/feat1.shtml)
2. Soergel, D. "The Rise of Ontologies or the Reinvention of Classification." JASIS, 50(12), 1999. pp. 1119–1120.
3. Koch, T. "Specification for Resource Description Methods. Part 3. The Role of Classification Schemes in Internet Resource Description and Discovery." A DESIRE deliverable. Section 3.7. (www.lub.lu.se/desire/radar/reports/D3.2.3/class_v10.html)
4. Shafer, K.E. "Automatic Subject Assignment via the Scorpion System." (www.oclc.org/oclc/research/publications/review96/scorpion.htm)
5. Larson, R.R. "Experiments in Automatic Library of Congress Classification." JASIS, 43(2), 1992. p. 147.
6. Morville, P. "Little Blue Folders." Strange Connections, July 10, 2000. (argus-acia.com/strange_connections/strange003.html)
7. SIG-IA discussion list. Weeks of 2/4/01 and 2/11/01. (www.asis.org/Conferences/Summit2000/Information_Architecture/listserv.html#archives)
8. Mulvany, N.C. Indexing Books. Chicago, University of Chicago Press, 1994. p. 46.
9. Chen, H., et al. "Automatic Thesaurus Generation for an Electronic Community System." JASIS, 46(3), 1995. pp. 175–193.
10. Rosenfeld, L. "Looking for Metadata in All the Wrong Places: Why a Controlled Vocabulary or Thesaurus is in Your Future." WebReference, February 2, 2001. (www.webreference.com/authoring/design/information/cv/index.html)

COPYRIGHT 2001, ARGUS ASSOCIATES, INC. ALL RIGHTS RESERVED.

About the Author

Kat Hagedorn (kat@argus-inc.com) joined Argus Associates as an Information Architect in 1998. Her background in biological sciences and information science (a B.S. from Cornell University and an M.S. from the University of Michigan School of Information) brings an added dimension to her expertise in designing organizational systems for clients.

Kat's previous publications include "Bottom-up Information Architecture: Leveraging Metadata" in SLA's Content Management book (1998) and the "Information Architecture Glossary" for the ACIA (2000). She has spoken in several venues, including at the American Society of Indexers Annual Meeting in May 2000.

At Argus, Kat has provided information architecture consulting services to a variety of clients including 3Com, Applied Materials, AT&T, Corning, Procter & Gamble, Square D and The Weather Channel.

COPYRIGHT 2001, ARGUS ASSOCIATES, INC. ALL RIGHTS RESERVED.

Pergamon

Government Information Quarterly 18 (2001) 195–208

Government Information Quarterly

The *Find-It! Illinois* controlled vocabulary: Improving access to government information through the Jessica subject tree

Anne Craig, Associate Director* ,
Suzanne Schriar, Digital Access Coordinator

Illinois State Library, Springfield, Illinois 62701-1796, USA

Abstract

The need to improve information access on the Web has resulted in Illinois' implementation of lexicographer Dr. Jessica Milstead's subject tree for the *Find-It! Illinois* Program. In 1999, when the Illinois State Library joined four other states in implementing a state Government Information Locator Service (GILS) project, developing a controlled vocabulary became an essential component for maximizing retrieval of government information. Furthermore, application of library cataloging tools such as the *Library of Congress Subject Headings (LCSH)* is insufficient for online retrieval. An analysis of the structure and content of Dr. Milstead's subject tree reveals the importance of new tools for improving online access methods. Illinois' implementation of Dr. Milstead's subject tree exposed the interest for nationwide application. The Illinois subject tree has been named the "Jessica Tree" to convey its expanded utility. The national adoption of a controlled vocabulary for retrieving state government information online will require collaboration among all states, so that the vision of a *Find-It! America* can be actualized. © 2001 Elsevier Science Inc. All rights reserved.

1. A context for developing and implementing an Illinois controlled vocabulary

The vast wealth of Illinois government resources available on the World Wide Web has made the Illinois State Library (ISL) view the application of a controlled vocabulary as essential for maximizing relevant resource discovery. Because the value of information is inherent in its utility, providing the best possible access to information remains the ultimate

* Corresponding author. Tel.: +1-217-795-5607.
E-mail addresses: acraig@ilsos.net (A. Craig), sschriar@ilsos.net (S. Schriar).

0740-624X/01/$ – see front matter © 2001 Elsevier Science Inc. All rights reserved.
PII: S0740-624X(01)00077-6

mission of the ISL. During a time characterized by the proliferation of information, abandoning controlled subject access in favor of imperfect keyword search strategies seems irresponsible and deceptively simplistic. Furthermore, the utility of established library classification schemes focused primarily on shelf arrangement of analog materials does not transfer to the Web.

Information access methods applied to a virtual medium requires new tools and standards for improving online browsing and retrieval capabilities. With a heightened awareness of the need for better retrieval mechanisms for government resources, the ISL embarked on the ambitious initiative to implement a nationally accepted subject tree for maximizing access to state government information.

1.1. IMLS grant one: "exporting Washington State's GILS" (October 1998-September 1999)

In early 1999, the ISL became one of five partners in an Institute for Museum and Library Services (IMLS) National Leadership Grant awarded to the Washington State Library (WSL). This IMLS grant project entitled "Exporting Washington State's GILS," provided a model for four state libraries to replicate a Government Information Locator Service (GILS) project based on the highly successful WSL GILS initiative (Washington State Library, 1998). The five participating state libraries, which included Illinois, New Hampshire, New Mexico, Oregon, and Washington, initially implemented the GILS project subject trees that were developed by Oregon and Washington (Oregon State Library, ND; Washington State Library, 1997).

Careful examination of the Oregon and Washington trees revealed that many subject categories were not uniformly transferable to disparate geographic regions of the United States. For example, the Oregon subject tree has the heading "Industry and Technology," with a subheading "Timber, lumber and logging." The main heading "Laws and Regulations," has subheadings for "Oregon administrative rules" and "Oregon revised statutes." Likewise, the Washington State subject tree shows the subheadings "Revised Code of Washington" and "Washington Administrative Code" directly under the broad category of "Laws and Regulations." There are numerous other examples, including subheadings for volcanoes, oceans, and salmon. Such state specificity in these subject trees made them less useful across state and geographic boundaries.

The Web has redefined society as a global community, whereby information needs are wider in scope and user expectations are more demanding than in any other decade. Additionally, the incalculable volume of information available has presented obstacles to access, highlighted by the limitations of our current information retrieval tools. As interstate searching becomes a possibility due to the availability of state government information on the Web, the demands for interoperable retrieval mechanisms have grown proportionately. ISL survey results indicate that users value information on states' activities, especially for the purposes of public policy and litigation development. "What are other states doing about..?" is the most frequently asked type of query from ISL users (Craig, 1998).

To meet these information needs, customizable yet uniform search strategies must be developed. Effective subject searching will depend on semantic interoperability, and a

standard controlled vocabulary can assist progress towards such interoperability. The challenge in developing a standard controlled vocabulary is preserving specificity (e.g., "salmon" in Oregon) within a universal context (e.g., "transportation"). The utility of a single subject tree is, therefore, dependent upon the application of broad, uniform categories in combination with regionally specific subcategories that can be displayed where appropriate and suppressed when not relevant. Knowing that the benefit of shared subject access was dependent on customization and ease of use at the regional level, the only way to implement a single subject tree was to develop a new one in collaboration with all participating GILS project state libraries.

1.2. IMLS grant two: "Find-It! Illinois" *(October 1999-September 2001)*

In Illinois, another impetus for revising the Oregon and Washington subject trees was the evolution of the *Find-It! Illinois* Program from a GILS project to a portal for all state library Web–based resources. The Illinois GILS Program site provides a searchable database that indexes state government Web sites. A state subject tree had to include categories broad enough not only to apply to the GILS database, but also to different types of *Find-It! Illinois* databases (Illinois State Library, 1999a). The *Find-It! Illinois* portal provides access to the following databases:

1. The Illinois Government Information site (IGI): which includes indexed state government Web sites (the ISL's GILS project)
2. The Illinois Digital Archives (IDA): which includes digital images and their associated metadata
3. The Virtual Illinois Catalog (VIC): which includes over 800 Illinois libraries' bibliographic records and holdings via a Z39.50/http search interface

In September 1999, the ISL received a two-year Institute of Museum and Library Services (IMLS) National Leadership Grant to further develop *Find-It! Illinois* as a statewide virtual library that would "serve to educate as well as improve direct, unmediated information access" (Illinois State Library, 1999b). This grant project builds on the strong foundation provided by the ISL's participation in Washington's IMLS project. Further, the ISL plays an important role in educating Illinois citizens and assisting them in finding and accessing important information. The ISL program staff also recognized that increasingly, electronic state government information, including digital images, provides an important supplement to print materials as a core educational medium. An integrated approach to various resources through *Find-It! Illinois* is a critical component in better citizen access. Ultimately, a shared subject tree was needed, not only to facilitate interstate subject access, but also to provide linkages to the databases accessed through the ISL *Find-It! Illinois* portal.

1.3. The three major components of Find-It! Illinois: *IGI, IDA, and VIC*

A major step in actualizing the goal of meeting the public's need for government information was the development of IGI, the Illinois GILS site. IGI is a search site with matching, subcategory, and fielded searching options available. It searches current State of

Illinois Web documents. Public information that is created or collected by all state agencies can be located through IGI. This service is extremely beneficial to all citizens, as it provides a single access point for government information from all Illinois State agencies.

The IGI resource resulted from the ISL's participation in the Washington State's 1998 IMLS project. For detailed information regarding Illinois' GILS Project implementation see <http://www.finditillinois.org/metadata/webmasters.htm>.

The foundation of ISL's GILS is metadata. Table 1 shows the metadata elements used by the state agency Web content creators for the purpose of identifying and describing their documents in the GILS implementation. The various sections of Table 1 list the types of data captured about each resource. In particular, Section 3 shows a sample of the controlled vocabulary. A drop-down menu is displayed with all the categories from the Jessica Tree (Illinois State Library, 2000). Web content creators select as many subject categories as are relevant. The selection of headings via displayed choices alleviates the chance of concealing resources through typographical or other manual errors.

The *Find-It! Illinois* Metadata Generator (*Find-It! Illinois*, 1999c) provides a template that allows Web content creators to add information easily. Resource descriptions supplied in the template are automatically translated into the appropriate metatdata elements. Once the fill-in, Web–based form is complete, the Webmaster can do a simple copy from the Generator and insert the HTML code into the head of the his document.

Overall, the impact of Web expansion has greatly exposed citizen demand for government information. Ultimately, the combination of the GILS metadata element set and the installation of GILS-compliant servers, will help to achieve reliable retrieval results among multistate and multiplatform systems.

Improving subject access has been a primary objective of the *Find-It! Illinois* Program, but it has not been the only one. Adding digitized documents from the ISL collections to the Illinois Digital Archive (IDA) also plays an important role in the development of the program.

IDA, which contains text as well as image files, is designed to be the electronic repository of the ISL's digital collections, as well as the digitized collections of other libraries and cultural institutions in the state. Currently, the material being digitized includes the ISL's complete collection of the *Illinois Blue Book* (1900–1999/2000) and the *Handbook of Illinois Government* (1970–1999/2000). Other items being digitized include the *Illinois State Academy of Science Transactions* (1908–1999); *Laws of Illinois* (1940–1998); and the *Illinois Senate Journal* (1978–1998). Information about IDA is accessible from the *Find-It! Illinois* portal: <http://www.finditillinois.org/ida.>

In addition to the material from the ISL collection that is being digitized on an ongoing basis, IDA will be augmented by digitized collections from other libraries, especially those that received digital imaging grants from the ISL. Subject areas in IDA (see Table 2) will tell the history of Illinois. Specifically, some of the target subject areas are the b icentennial of Illinois' statehood (2018); the bicentennial of Lincoln's birth (2009); the Nauvoo Mormon uprising (1844); WPA projects (Chicago was the Region 10 Headquarters); coal mining (especially labor disputes and mine disasters); and Carnegie libraries.

As the Web is ubiquitous, so are ISL users; therefore, it is incorrect to assume that the scope of their information needs is limited to government information. Clearly, technology

Table 1
Find-It: Illinois metadata elements for indexing state government information for igi
Section 1: Basic properties about your site

Element	Description and example
A. Site title	The officially assigned title of your document/page *<meta name = "siteTitle" content = "Illinois State Library Homepage">*
B. Keywords	Includes any terms, acronyms and/or phrases that users may apply to search for information *<meta name = "keywords" content = "Junior League of Springfield>* *<meta name = "keywords" content = "The Springfield Junior League">*
C. Description	A narrative that summarizes the content and purpose of your document/page. *<meta name = "description" CONTENT = "The Department of Business Services is a centralized repository for business filings in six areas: Corporations (For profit and NOT for profit), Limited Liability . . .>*
D. Date created	The date your document was created. *<meta name = "createDate" content = "1997/07/18">*
E. Date last modified	Used each time you change any content of the document. *<meta name = "dateofLastModification" content = "1998/03/01">*
F. Time period	Timeframe covered by the document's content *<meta name = "timePeriodTextual" content = "Statistics collected for FY 1993–1995">*
G. Medium	Website, CD-ROM, Database, Microfilm, Paper, Photograph, Sound Recording or Image. *<meta name = "medium" content = "website">*
H. Site to be retained for permanent access?	"Yes" is the default, since most documents will be public *<meta name = "PermanentPublicAccess" content = "Yes">*
I. Unique control ID or number?	Used for an official control ID assigned to your document (if applicable) *<meta name = "originalControlIdentifier" content = "1040A">*
J. Agency program	Official name for the agency program for your document. *<meta name = "agencyProgram" content = "Crime Victims Compensation">*
K. Related sources	Allows a direct link to a point of reference to other helpful information on the Web *<meta name = "linkage"* content = http://www.revenue.state.il.us>
L. Language	English is the default *<meta name = "language" content = "EN">*
M. Government type	To indicate either state or local document *<meta name = "language" content = "state">*

Section 2: Contact information about your site

Element	Example
A. Contact Name	*<meta name = "contactName" content = "Bonnie Matheis">*
B. Organization	*<meta name = "contactOrganization" content = "Illinois State Library">*
C. Contact Address line 1 & line 2	*<meta name = "contactStreetAddress1" content = "300 South Second Street">* *<meta name = "contactStreetAddress2" content = "Third Floor LAT">*
D. City	*<meta name = "contactCity" content = "Springfield">*
E. Zip	*<meta name = "contactZipcode" content = "62701-1796">*
F. Email	*<meta name = "contactNetworkAddress" content = bmatheis@ilsos.net>*
G. Phone	*<meta name = "contactPhoneNumber" content = "(217) 558-2065">*
H. Fax	*<meta name = "contactFaxNumber" content = "(217) 557-6737">*

Table 1 (continued)

Section 3: Subject analysis
Agriculture and food production
Agriculture and food production-Agricultural finance
Agriculture and food production-Agricultural statistics
Agriculture and food production-Agricultural pages for kids
Agriculture and food production-Aquaculture
Agriculture and food production-Crops
Agriculture and food production-Fisheries
Agriculture and food production-Irrigation
Agriculture and food production-Livestock
Agriculture and food production-Pesticides
Agriculture and food production-Soil erosion
Business and industry
Business and industry-Banking
Business and industry-Business and professional licenses
Business and industry-Business and professional licenses-Business licenses

has redefined the concept of library collections to extend beyond the physical boundaries of the library, thereby supplementing ownership with access to digital resources.

To expand the scope of access to include bibliographic records and associated holdings information, the ISL developed the Virtual Illinois Catalog (VIC) to provide a single Web interface to consortial library catalogs within the state. More precisely, VIC provides bibliographic and holdings records of the 800+ library collections contained within the member libraries of the twelve subregional library systems throughout Illinois. The virtual catalog uses the Z39.50 information retrieval protocol to provide users with the capability of searching multiple library catalogs concurrently. A detailed description of VIC, including a summary of its history, scope of participating libraries, frequently asked questions and searching hints, is available from the *Find-It! Illinois* portal: <http://www.vic.lib.il.us>.

The information retrieval model underlying *Find-It! Illinois* is designed to eventually allow a user to execute a single query across multiple databases simultaneously. Thus, a user will be able to search across the GILS database (i.e., the IGI site), the Illinois Digital Archives, and the Virtual Illinois Catalog with one search.

This model, that is, executing a single query across multiple databases, presents a

Table 2
Jessica Tree top level subject categories

Agriculture & food production	Land use & construction
Business & Industry	Law enforcement
Consumer protection & safety	Laws & regulations
Education & libraries	Natural resources
Forms	Recreation and tourism
Government	Society & culture
Health & social services	Utilities & transportation
Information & telecommunications	Voting & elections
Kids pages	

challenge to the *Find-It! Illinois* Program. Initial examination of the databases accessible via the *Find-It! Illinois* site (i.e., IGI, IDA, and VIC) reveals differences in the access tools and the underlying structure of the data in each. Subject access across the various databases is of particular concern.

Subject access to bibliographic records in VIC servers is primarily limited to the *Library of Congress Subject Headings* (LCSH; Library of Congress, 2000), whereas access in the IGI index is based on the *Find-It! Illinois* controlled vocabulary, the Jessica Tree. IDA uses a combination of both the LCSH and the Jessica Tree. Semantic interoperability across *Find-It!* databases will require a complex crosswalk (currently in development at the ISL) between LCSH terms present in IDA and VIC, and the Jessica Tree terms present in IGI and IDA. Currently, there are numerous cross-references from popularized terms to LCSH headings within the Jessica Tree. There are also cross-reference from LCSH headings to popular ones. The work that remains to be completed can be described as developing crosswalks among all headings in the Jessica Tree to specific headings in LCSH.

In its IMLS grant application, the ISL maintained that these valuable initiatives, namely IGI, IDA, and VIC, are crucial to the overall mission of the ISL. Ultimately, the ISL charge to offer one central site for searching, selecting, and maintaining electronic resources, resulted in the *Find-It! Illinois* portal. Also informing for the design of the portal is the ISL's goal to meet the challenge of providing coherent and integrated subject access across the different resources. This goal required the development and deployment of an Illinois controlled vocabulary.

2. The case for controlled vocabulary

The impetus to develop and apply controlled vocabulary for searching state documents online gained momentum from the imperfection of keyword–driven search engines to retrieve complete and uniformly relevant information. Furthermore, the overwhelming number of items retrieved with keyword searches has emphasized the need for more control and organization of information on the Web. While users rely heavily on keyword searching to yield results, relevant resources are often missed. Furthermore, the lack of consistent content among metadata records emphasizes a strong need to impose more rigorous control over how we represent the subject content of Web documents.

Eliminating natural language or keyword searching as a tool for information retrieval is certainly not an objective at the ISL. Rather, supplementing random keyword access with established, authorized topical headings that represent the contents of resources provides the basis for organizing and locating electronically existing resources. Use of natural language, while providing the basis for communication, does not provide the best means of information retrieval. Additionally, the vast range of synonyms and varying forms of words render keyword searching unreliable. Ultimately, controlled vocabulary makes predictable search results possible. The structural design of a controlled vocabulary with its elaborate interrelationships between terms provides flexibility in assigning appropriate terms to documents and assists users in determining appropriate controlled terms for their queries.

Controlled vocabulary maps natural language terms to standardized and authorized terms

and thereby chooses one term among many to represent a topic or concept." In addition, a controlled vocabulary shows relationships between separate yet similar subjects that are listed under different authorized headings. Because language provides variant ways to express the same subject or concept, there remains a need to link and map synonyms, plural endings, and alternative spellings to a single heading.

Fundamental relationships between terms (e.g., synonyms, homographs, and hierarchy) reflected in a controlled vocabulary allow a natural language term to be linked to the controlled term to ensure a successful match. The beauty of these interrelationships is evidenced in that the ease of keyword searching is maintained. Controlled vocabulary links the searcher's language with the Web content creator's language to make an exact match possible. The relationships are constructed as cross-references that are not visible to the searcher. The Web content creator selects from a drop-down listing of the subject tree the term most appropriate for subject representation. The searcher can either drill down within the subject hierarchy to select the authorized heading or search by keywords. When keywords are searched, nonstandard terms are automatically linked to the authorized ones.

To some degree, metadata records, which provide structured descriptions about data within the framework of a standard set of metadata elements, improve keyword access to electronic resources. For example, a metadata record can record information in a structured manner about a title of a resource, the date of publication, the author, and so forth A metadata scheme provides a structured framework for describing electronic resources, but it does not necessarily prescribe the data content for each of the metadata elements. Requiring the use of a structured, controlled vocabulary for data content in, for example, the "subject" metadata element becomes a foundation to improve information retrieval. In practice, subject terms must be embedded in metadata records to affect uniform retrieval. Metatdata element sets merely provide a framework for organizing various virtual containers for descriptive elements of the resource. Whereas metadata provide virtual containers for descriptive date, controlled vocabulary standardizes the content within the containers. To reiterate, as Milstead and Feldman (1999) point out:

Metadata standards... are limited almost entirely to formats and frameworks, meaning standardization of the tags and of packages for them.... Metatdata formats are likely to provide a means to specify the authority used for the content of the tag, but they are designed as packages and it is up to the user to design the content of the package.

Furthermore, the metadata content necessary for subject access requires a complex standardization of terms and interrelationships. The insufficiency of natural language to furnish complete and reliable access to information on the Web has exposed the need for, and consequently the opportunity to develop more effective methods of organizing our vast, virtual knowledge base.

As evidenced by the *Library of Congress Subject Heading (LCSH)* list and other subject heading schemes and thesauri, control and standardization does not imply a static and unchanging vocabulary. Rather, ongoing revisions to establish new headings and to eliminate old ones are crucial to preserving currency and ultimately authority and utility of a controlled vocabulary. Application of a controlled vocabulary that is not periodically examined for necessary modifications outlives its usefulness. As librarians and agents for resource sharing

and information retrieval, our mission must be to ensure that information is uncovered, not suppressed by the application of our access tools.

Controlled vocabulary has a vital role to play in assisting subject searching and access in the electronic environment. The question is, "which controlled vocabulary will work best for a particular application?" LCSH's broad scope covers all subject areas, contains a rich structure of cross-references, and facilitates subject browsing, but there are some serious limitations in its use for Web resources. Among these limitations is the fact that the complex rules for assigning LC subject headings require much cataloging expertise. Web content creators are not typically information specialists or librarians. To expect that they become knowledgeable and willing to apply cumbersome and complex access tools designed by and for librarians is unrealistic. Furthermore, LCSH is incompatible in design and syntax with most other controlled vocabularies; and therefore, does not foster interoperability. After an analysis of existing subject heading lists and thesauri, the ISL made the decision to construct a flexible controlled vocabulary for its use in *Find-It! Illinois.*

3. The making of the Jessica tree

The ISL contracted with Dr. Milstead in late 1999 to write a subject tree that would facilitate searching other states' GILS sites as well as provide interoperability across the *Find-It! Illinois* databases. Dr. Milstead constructed the tree by examining GILS and library-based Web subject representation schemes used in many different states. Three primary schemes emerged in her investigations: Minnesota's *Legislative Indexing Vocabulary* (Minnesota Office of Technology, ND), the Oregon and Washington subject trees, and the *Texas Records and Information Locator List of Subject Headings* (Texas State Library and Archives Commission, 2001). Table 3 lists the charge from the ISL to Dr. Milstead's charge for a hierarchical subject tree.

The Jessica Tree is based upon a networked arrangement of terms, where categories are repeated in multiple places within the tree. Within this network of browsable categories, there is also a hierarchical dimension, as revealed by general categories at the top level to most specific at the lowest level. This arrangement provides an optimal means of viewing the complete range of narrower terms under a specific browsable category. There are a total of seventeen top–level categories (see Table 4). Under each top-level category is an elaborate structure of narrower terms ranging in depth from second level subheadings to, in some cases, fifth and even sixth levels of narrower headings.

The purpose of the hierarchical structure is to allow the user to either select a topic at the broadest level to increase the number of hits, or drill down to a more specific level to facilitate greater precision of search results. One example, showing the depth of narrower and broader subject terms under a single heading is found under the heading "Utilities and transportation." "Transportation" is found at the second level. At the third level are thirteen types of transportation modes ranging from air to water vehicles. At the forth level are headings found for types of motor vehicles, including "Automobiles," "Buses," "Motorcycles," and "Trucks" (see Table 5).

By contrast, an alphabetical arrangement, as found in the LCSH, merely implies an

Table 3
Example of Jessica Tree hierarchy

TOP LEVEL CATEGORY	SECOND LEVEL CATEGORIES UNDER *"UTILITIES & TRANSPORTATION"*	THIRD LEVEL CATEGORIES UNDER *"TRANSPORTATION"*
Utilities & transportation	Drinking water	Air transportation
	Electric utilities	Airports
	Energy resources	Automobiles
	Gas utilities	Bridges and tunnels
	Railroads	Buses
	Roads and highways	Mass transit
	Sewage treatment and disposal	*Motor vehicles*
	Telecommunications	
	Traffic safety	
	Transportation	
	Trucks	
	Utilities and transportation	
	Pages for kids	
	Water transportation	

FOURTH LEVEL CATEGORIES UNDER *"MOTOR VEHICLES"*	FIFTH LEVEL CATEGORIES UNDER *"AUTOMOBILES"*	SIXTH LEVEL CATEGORIES UNDER *"DRIVERS LICENSES"*
Automobiles	*Drivers licenses*	License application forms
Buses	Sport utility vehicles	
Motorcycles	Vehicle registration forms	
Off road vehicles		
Trucks		
Vehicle registration and titles		

underlying hierarchy based on references to broader, narrower, and related terms. However, the implicit hierarchy in LCSH does not facilitate browsing in an online environment. Essentially, the linear structure of subheadings does not easily allow the user to select or drill down to increasingly narrower topics. The translation of subheadings into more specific categories is an important utility of online browsing, and therefore was an important requirement for the Jessica Tree.

Another helpful component of the subject tree is the placement of numerous topic terms embedded under more than a single category. Also, in some cases, terms are repeated under a single subject category, in varying levels of depth within the subject tree in order to maximize ease of access to particular topics. The heading ***"Vehicle registration forms"*** provides one illustration of multiple occurrences of a topic within the hierarchy. ***"Vehicle registration forms"*** appears numerous times within the Jessica Tree. These include a second level heading under the top category "FORMS," as well as additional occurrences under the heading "Transportation" which is found under the top-level category "UTILITIES AND TRANSPORTATION". ***"Vehicle registration forms"*** shows up under the following subheadings: Automobiles; Motor vehicles; Buses; Motorcycles; and Trucks. Additionally, ***"Vehicle registration forms"*** appears under

"Drivers licenses," which occurs under each subheading listed above. Essentially, many paths lead to the desired information.

3.1. Controlled vocabulary and searching Illinois government information

The architecture of the *Find-It!* Illinois Government Information Program (IGI, the ISL's GILS project) supports a variety of different approaches to search for each Web resource:

- Subject metadata element (controlled vocabulary selected from the Jessica Tree display in the metadata generator)
- Keyword metadata element (noncontrolled terms supplied by the Web content creators)
- Keyword searching of entire resource (where no metadata record exists)

The application of the Jessica Tree in the IGI search facility expands the original search terms to include equivalent and related terms (i.e., see and see also/broader and narrower terms).

Currently, IGI usage logs reveal keyword access to be the most commonly used retrieval method by the general public. Future enhancements, however, will include natural language processing, which results in answers to searches expressed in the form of questions, for example, "When did Illinois become a state?

Optimally, controlled vocabulary that is employed with the option to use free text or natural language is important for formulating a search strategy that ensures the best retrieval results. In fact, keyword access is useful for supplementing controlled vocabulary for retrieval of full-text items and to increase the hit rate.

4. From the state to the nation: collaboration on a national subject tree for government information

In October 1999, IMLS funded the Washington State Library's proposal, "Metadata Tools for State Collaboration," (Washington State Library, 1999) to promote the development of a nationwide subject tree Washington, in turn, contracted with the ISL to continue work-in-progress on the Jessica Tree.

In January 2001, Kelly Meier, a freelance library consultant hired by the ISL, began intensive interviews with each state toward meeting the goals of:

(a) identifying the obstacles endemic to national adoption of the tree,
(b) heightening awareness of the project, and
(c) designing implementation strategies.

Meier's distributed a questionnaire to electronic mail lists and state libraries in January 2001 that asked the following:

1. How electronic government information is categorized
2. Which metadata element sets are used
3. Which subject schemas are used to navigate GILS projects

4. Comments on crosswalks
5. Comments on the desirability of a shared subject tree

The survey is available online (Meier, 2001).

In March 2001, a nationwide conference on the implementation of the Jessica Tree occurred in conjunction with the Third Annual State Government GILS Conference in Springfield, Illinois. Ms. Meier presented the results of her survey and work has continued to expand the utility of the Jessica Tree for other states.

4.1. Find-it! America

As state libraries' GILS projects have evolved over the last three years, interest in a single unified *Find-It! America* project is emerging. A *Find-It! America* initiative, building on the work begun by the WSL, could result in the implementation of consortial methodologies such as technology architecture (including the adoption of the Jessica Tree), new state project start-up, marketing, purchases, and public policy development. The GILS Conference in Springfield provided an arena for beginning the work toward the common goal of establishing a *Find-It! America* consortium.

5. Conclusion

All state libraries participating in the third IMLS project to develop a shared subject tree and to ultimately establish a *Find-It! America* appreciate fully the elusiveness of an accepted shared vocabulary scheme, but nonetheless are committed to the success of the project. In the discipline of information science, the value of a universally applied controlled vocabulary scheme has been acknowledged as paramount for over 100 years, when in 1898, the Library of Congress developed LCSH.

The state library, for the majority of states, is the logical headquarters for a GILS project. Most states have some sort of documents depository program. Expansion to electronic state government information under the authority of the state library is reasonable. Additionally, a number of states have now adopted legislation specifying the state library as the provider of electronic government information.

State libraries can, in the words of Nancy Zussy, Director of the WSL, "... offer government agencies a ready-made cooperative network of information centers, located in virtually all communities... " She continues, "Policymakers and the general public are coming to recognize the content management expertise of librarians, skills that transcend information formats and technology of various kinds," (Zussy, 2000).

At present, state government organizations routinely serve electronic information on the Web; therefore, state libraries play a role in providing access to electronic government information to citizens when value is added through the use of tools described in this paper. Specifically, this value can be realized most dramatically in the acceptance and application of a single vocabulary for state government resource description and subject representation. This controlled vocabulary provides the key to extending information retrieval across state lines and among disparate state library electronic collections.

Many metadata element sets, that is, the tags themselves, have been mapped to one another. In fact, the Extensible Markup Language (XML) (OASIS, 2001) and the Resource Description Framework (RDF) (W3C, 1999) standards may provide the automated framework to organize and translate the labyrinth of these sets. However, translation among vocabularies (i.e., semantic interoperability or semantic mapping) is inexact and provides the reason for the establishment of a standard scheme. The content of the metadata elements themselves, especially the element that contains subject terms, is where the states' GILS projects should be most intensely focused.

Successful implementation of an interstate searching capability for government information would fulfill not only a primary user demand, but would also serve as a subsequent step toward a nationwide consortium, thereby creating unprecedented state library cooperation on a programmatic level. Interstate searching capability is yet another step in the evolution from the original Washington State's GILS project. Although a formal consortium may not be necessary to further the goals of the *Find-It! America* vision, cooperative efforts may be more readily adopted and legitimized by policymakers if their origin is an official body.

In general, the transition of state library government documents programs from hard copy to electronic based has been inconsistent and reactive. The emergence of government information on the Web has left state library administrators with the dilemma of supporting paper-based and electronic programs simultaneously. *Find-It! America* will provide a national strategy for revolutionizing cost-effective, replicable state library program development. Ultimately, the implementation of the *Find-It! America* concept, on the virtual horizon, promises to create not only an information access framework for the global community of state government information seekers, but also a framework for information specialists to manage and provide access to critical digital resources.

References

Craig, A. (1998). Illinois State Library. *Illinois State Library User Survey* (unpublished).

Illinois State Library. (2000) *The Jessica Tree.* Available URL: http://www.finditillinois.org/metadata/webmasters.htm. Accessed February 23, 2001.

Illinois State Library. (1999a). *Find-It! Illinois* (includes access to IGI, IDA, and VIC). Available URL: http://finditillinois.org. Accessed March 23, 2001.

Illinois State Library. (1999b). *Find-It! Illinois IMLS National Leadership Grant.* Available URL: http://www.cyberdriveillinois.com/library/isl/lat/lat imls grant.html. Accessed March 23, 2001.

Illinois State Library. (1999c). *Find-It! Illinois Metadata Generator.* Available URL: http://www.finditillinois.org/metadata/index.html. Accessed March 23, 2001.

Library of Congress. (2000). *Library of Congress Subject Headings* (23rd ed.). Washington, DC: Library of Congress.

Meier, K. (2001). *Digital Government Information Survey.* Available URL: http://www.fidocat.com/gils/survey.htm. Accessed February 23, 2001.

Milstead, J., & Feldman, S. (1999, January-February). "Metadata: Cataloging by Any Other Name." *Online, 23.* Available URL: http://www.onlineinc.com/onlinemag/metadata. Accessed February 23, 2001.

Minnesota Office of Technology. North Star, Minnesota State Government Online (N.D.). (2001). *Legislative Indexing Vocabulary.* Available URL: http://bridges.state.mn.us/servlet/lexico?pub-22&frames&LIV-A&. Accessed on March 23, 2001

OASIS. (2001). Organization for the Advancement of Structured Information Standards. *The XML Cover Pages.* Available URL: http://xml.coverpages.org/xml.html. Accessed March, 23, 2001.

Oregon State Library (N.D.). (2001). *Compass: Category: FindOr category tree: ROOT.* Available URL: http://www.state.or.us:1080/compass. Accessed March 6, 2001.

Texas State Library and Archives Commission. (2001). *TRAIL (Texas Records and Information Locator) List of Subject Headings.* Available URL: http://www.tsl.state.tx.us/trail/manualappendix3.html. Accessed March 23, 2001.

Washington State Library. (1999). *Metadata Tools for State Collaboration.* Available URL: http://www.statelib.wa.gov/projects/imls/2001/index.htm. Accessed March 23, 2001.

Washington State Library. (1998). *Exporting Washington State's GILS.* Available URL: http://www.statelib.wa.gov/projects/imls/webpage.html. Accessed March 23, 2001.

Washington State Library. (1997). WSL: Washington GILS subject tree. Available URL: http://www.statelib.wa.gov/info rscrs/dbs tools/find-it/gilstree.htm. Accessed March 6, 2001.

World Wide Web Consortium (W3C). (1999). *Resource Description Framework (RDF) Model and Syntax Specification.* Available URL: http://www.w3.org/TR/1999/REC-rdf-syntax-19990222. Accessed March 23, 2001.

Zussy, N. (2000, May). "The Birth and Development of Find-It!: Washington State's Government Information Locator Service." *First Monday,* 5. Available URL: http://www.firstmonday.dk/issues/issue5 6/zussy/. Accessed February 23, 2001.

ELSEVIER

Journal of Hazardous Materials 61 (1998) 291–297

JOURNAL OF HAZARDOUS MATERIALS

An outline of the approach being used in developing the OECD Dictionary/Thesaurus of 'risk assessment' terminology

Isadore Rosenthal *, Ben Ale, Lyse D. Helsing

The Risk Management and Decision Processes Center, The Wharton School of the University of Pennsylvania, 1325 Steinberg Hall-Dietrich Hall, Philadelphia, PA 19104-6366, USA

Abstract

A system was developed to capture and make available the definitions of terms and the approaches and methodologies in the risk assessment process of the member-countries of the OECD. The system uses the principles of artificial intelligence. The system is running on the Internet. © 1998 Published by Elsevier Science B.V. All rights reserved.

Keywords: OECD; Communication; Internet; System

1. Introduction

The OECD Expert Group on Chemical Accidents set out the Project's goals rather clearly:

> "In summary, the workshop participants concluded that standardization of the risk assessment process, and approaches/methodologies used in each step of the process, is neither desirable nor feasible. Nonetheless, enhancing the mutual understanding of risk assessment in the context of chemical accidents, can be furthered by, e.g., efforts to map out the steps in the risk assessment process and the approaches/methodologies used therein, and an elaboration of the indicators influencing choices of particular approaches/methodologies. It must be emphasized that this is not intended to direct, still less to prescribe, a particular approach.

* Corresponding author.

0304-3894/98/$19.00 © 1998 Published by Elsevier Science B.V. All rights reserved.
PII S0304-3894(98)00135-6

The objective of these efforts should be to help stakeholders see more clearly the range of possibilities, and to assist them in decisions which only they can make.'' [1].

The project goals set out in this statement can be framed as follows:

- Illuminate the sense of the various approaches to risk assessment described in the literature.
- Promote understanding of the commonalities (and differences) among the various approaches to risk assessment.
- Facilitate communication of such commonalities (and differences) between different risk cultures and languages.

There is a need to constantly remind oneself that the project's goal is to provide 'products' that promote understanding and communication among the project's stakeholders, and not to set up a new terminology aimed at replacing or rationalizing present systems. Over time, stakeholders will likely take actions on their own that will lead to convergence, but that is not the project's charge.

To say that achieving this project's goals is an ambitious undertaking is somewhat of an understatement. The difficulty is compounded by the fact that some of our stakeholders who claim to do risk assessment do not, and others deny doing it, and in fact do.

The Thesaurus must therefore allow the 'owners' of an item they are entering to find terms in the locations they believe they belong, even if our generic representation or logic indicates they do not belong there. 'Official' and technical UK, French, Netherlands and American usages each have somewhat different logics. The Thesaurus makes an attempt to allow for the collection and illumination of terminology and concepts in all of these logics.

The measure of success will be whether or not stakeholders find the Thesaurus useful. To accomplish this, the Thesaurus must strike an appropriate balance between our stakeholders' needs for completeness and uniqueness, and their need for a Thesaurus that is not too technically complex for practical use.

Clearly, there will have to be trade-offs and the project's Steering Group will have to do a Pareto analysis and settle for meeting, say 95%, of our stakeholders needs for completeness and uniqueness in order to reduce technical complexity.

2. The conceptual approach used in developing the Thesaurus

Each item entered into the Thesaurus will have an identified 'owner', the 'client'. The client's intended meaning of their entry will be elicited by a questionnaire containing a series of structured, operationally phrased, questions organized in a hierarchical system. These questions will be designed to elicit the greatest level of informational detail embedded in the owner's meaning/understanding of the item being entered.

In some ways, the project's approach to the Thesaurus can be visualized as conceptually similar to the taxonomic approach used for the description, identification, naming of organisms, and their classification into hierarchical groups. Others have

viewed the Thesaurus as a 'translation Engine' which captures the intended meaning of a risk assessment item and 'translates' it into defined operational language.

The questionnaire is available as a document on the Internet. Clients convey the meaning they attach to the risk assessment content of an entry (definition, regulation, code, risk assessment case) by their responses to a hierarchically organized series of questions. The client's 'response' to each question is captured in a database linked to the client's plain language description of their entry.

Users will be able to query or search the Thesaurus on-line. Standardized reports giving the content of an entered item or comparing two or more entries will be available. Some capacity for Boolean searches of the Thesaurus database will also be available in the future.

3. Construction and nature of the hierarchial system

Three major tasks needed to be addressed in order to reduce the Thesaurus concepts outlined above to practice: (1) A suitable hierarchical system (set of questions) to elicit the needed inputs on an item from the client, i.e. a suitable taxonomic system. (2) An effective and efficient physical system for collecting and storing the inputs. (3) A system that allows the user to obtain desired Thesaurus outputs.

After examination and analysis of the manner in which the risk assessment process is described and categorized in the literature, a decision was taken on the broad categories that would be used to structure the hierarchical system for capturing inputs and furnishing outputs. The categories were chosen, to the extent practicable, to capture and mirror the present ways in which different risk cultures view the risk assessment process.

While the system was designed with terminology suited to describe accidental releases of chemicals from fixed facilities and the undesired outcomes from such releases, a reasonable effort was expended to develop categories that could be used in any other risk area.

Particular attention was paid to avoiding 'terms of art' such as dose, exposure, risk, etc. for obvious reasons. This, of course, led to increased verbosity which must be endured, though hopefully minimized. One cannot convey clear meaning using words whose meaning is equivocal.

The system is organized hierarchically in five levels. At the top level there are four broad Generic Elements. Each Generic Element has a varying number of sub-Elements, and each Sub-Element has a varying number of Terms. Each Term has is broken into categories containing a set of Descriptors which make up the lowest level of the system.

The terminology used to describe these hierarchical levels is as follows:

Element: A group of related, operationally defined, general concepts.

Sub-Element: One or more closely related, operationally defined, concepts embodied in the Element to which it relates.

Term: A single concept in the particular Sub-Element to which it relates.

Category: Examples that give general operational meaning to a concept and contain a set of related Descriptors.

Descriptor: Examples that give a specific operational meaning to a concept.

The Categories and Descriptors in the Thesaurus are only a set of examples selected because they were most likely to cover common usage. Clients can create their own Categories and Descriptors as required to capture their intended meanings or understanding of their entry, and the system will capture and display such entries. Over time, multiple entries of a new Category or Descriptor will be incorporated into revised versions of the Thesaurus, and unused ones will be dropped.

The number of sub-Elements and their Terms, Categories of Descriptors and Descriptors reflect both logic and system needs. They are set to promote ease and effectiveness in collecting the information needed to capture the intended meaning of the person submitting an item and to convey this meaning to users of the Thesaurus. While the Sub-Elements of Element I contain only one Term each, the sub-Elements of Element IV contain as many as four Terms.

The number of Elements, Sub-Elements, Terms, Categories and Descriptors is as follows: *Elements*: 4, *Sub-Elements*: 14, *Terms*: 19 *Categories*: 70, *Descriptors*: 368.

In order to give an impression of the terminology used in the system, the definition of the four terms is given here.

Generic Element I: Identification of sources with the potential to cause undesired outcomes to subjects of concern.

Generic Element II: Identification of possible sequences of events leading to loss of containment of the potential to cause undesired outcomes to a subject of concern resulting in its entry into a domain of the ecosystem. Estimation of possible distributions of both the released potential and the subjects of concern over time periods within compartments delimited by specified boundaries or end-points.

Generic Element III: Identification and description of how the specified undesired outcome is related to the intensity, time and mode of contact of a specified potential to cause the undesired outcome to the subject(s) of concern.

Element IV: Identification of the basis for estimating and expressing the likelihood that a specified undesired effect* will occur and description of the quality/uncertainty of such estimates; comparison of the estimates with relevant standards and guidelines and evaluation of the impact of specified alternative assumptions on the estimates.

As will be discussed later, the client has the ability to generate new 'Categories' and 'Descriptors', and insert them into the report on his entry. However, the number of Elements, Sub Elements and Terms are fixed.

The four Risk assessment Generic Elements are designed to capture only a segment of a continuum of activities that start with scoping the risk assessment, run through the risk assessment process and continue onto activities related to risk management, risk reduction, emergency response, risk communication, etc. Two additional Elements, O and V, are presented as simple text boxes. The purpose of these text boxes are to capture information that a client submitting an item feels is part of the risk assessment process not captured within the somewhat arbitrary boundaries used in the Thesaurus.

Element 0 is meant to capture aspects of the item being entered that the client believes should be included as part of the risk assessment process, and which generally precede Element I, e.g. scope and purpose of the Risk Assessment, stakeholder interactions re assumptions, etc.

Similarly, Element V covers aspects that generally follow Element IV, e.g. the Risk Assessment–Risk Management interface.

4. The computer system for collecting inputs and providing outputs

A large amount of complex information needs to be collected from and subsequently be made available to stakeholders who are widely distributed geographically in order to meet the project's objectives. Moreover, the Thesaurus needs to be capable of easy, quick and inexpensive updating, since its contents can be expected to change fairly rapidly over time. Ideally, it should also be available in a number of languages (commitment for a French version already exists).

Given these needs, the Thesaurus is being built as a computer document, easily accessible on the Internet. The 'input' questionnaire can be filled in by selection of appropriate 'boxes' on a HTML document. Instructions and clarification of terminology are available at each data entry point.

5. Entering information into the Thesaurus

The Thesaurus can handle the risk assessment content of four classes of items: Definitions, Laws and regulations, Specific risk assessment studies and Guidelines, policies, or codes.

The system presents the client with the opportunity of serially selecting applicable Elements, sub-Elements and ultimately any appropriate descriptors related to the item being entered. Provisions exist to collect reference information on each selected Descriptor, related 'Criteria' determining membership in the Descriptor class and 'Tools', (i.e. methodology, models, procedures etc.) associated with the selected descriptor. The client entry process has been summarized elsewhere [3].

5.1. Thesaurus outputs

Outputs will be furnished by essentially reversing the hierarchical system used for collecting inputs. Users will able to ask for the intended meaning of a definition or the risk assessment provisions of a regulation, guideline, or an actual risk assessment case entered by a specified agency, society or individual.

Three types of Thesaurus outputs are visualized:

A complete report on an entered item.

Comparisons of entries on the same entered item by different clients, comparisons of different entry items by the same client. These comparisons will be able to be made at a chosen level of depth, i.e. Elements, sub-Elements, terms, Category of Descriptor and Descriptors.

Reports identifying all entries that contain or refer to Elements, Sub-Elements, Terms, Category of descriptor or descriptors specified by the person making a search.

6. Discussion and conclusions

The initial test of the Thesaurus involved four countries. Each country had two participants enter the risk assessment provisions of the old 'Seveso' Directive and the country's laws implementing or corresponding to the Directive. This test revealed several deficiencies in both the ability of the system to capture a client's intended meaning (content) and the operation of the computer entry system [2].

The principle deficiency in the area of content arose in regard to interpreting and capturing explicit vs. implicit provisions of the document being entered. For example, a provision in a regulation that 'regulated facilities must notify all residents potentially affected by an accidental release' was interpreted by some clients as a risk assessment requirement for quantitative consequence analysis, although the particular regulation did not explicitly mandate such an analysis per se. Very many examples of this type of ambiguity were uncovered. This difficulty has been resolved by making system provisions allowing the client to differentiate between descriptors selected because they believed that the item being entered contained them explicitly or implicitly. Reports on entries will also allow for this type of differentiation. Limitations on the length of submitted papers does not permit a more detailed analysis of this trial in this paper.

A second trial of the Thesaurus based on the improved system started on July 1, 1997. This trial will involve entry of risk assessment guidance for facilities that store and process specified quantities of Chlorine and or LPG. In addition, each participant will enter their definitions for ten specified risk assessment terms, e.g. hazard, risk assessment, risk analysis, risk, exposure, etc. The results of this test will be available from the system via a greatly enhanced reporting/search system at the time the paper is being presented at this conference.

At this point in time, it is not possible to evaluate the utility of the OECD risk assessment Dictionary/Thesaurus to a wide audience of involved but non-technical specialists. However, it is already clear that the Thesaurus provides the technical community new insights into the structure and communication of the risk assessment process and the preparation and comparison of risk assessment studies and regulations.

Acknowledgements

This work was sponsored and supported in large part by the OECD Expert Group on Chemical Accidents. The authors also wish to acknowledge the considerable contributions made by members of the project's OECD Steering group, and significant contributions by members of an Ad Hoc industry group, composed of D. Ferguson of ICI, M. Jayjock and D. Hendershot of Rohm & Haas and A. Ignatowski, S. Schechter, J. Lynch and P. McNulty. Special acknowledgement must be made for C. Ross at the University of Pennsylvania for developing the Internet presentation with help from S. Kimbrough. Work supported in part under the U.S. Environmental Protection Agency's Cooperative Agreement 821947.

References

[1] Draft conclusions and recommendations concerning the sessions on risk assessment, OECD Workshop on Risk Assessment and Risk Communication in the Context of Accident Prevention, Preparedness and Response, Item 15, Paris, July, 1995.

[2] Second meeting of the OECD Steering Group on Risk Assessment in the Context of Accident Prevention, Emergency Preparedness and Response, Document 2, 14–15 February 1996, Washington, DC.

[3] A.J. Ignatowski, I. Rosenthal, L.D. Helsing, An Internet Thesaurus/Dictionary for Analyzing Risk Assessment Processes, Laws and Regulations, International Conference and Workshop on Risk Analysis in Process Safety, October 21–24, 1997, Atlanta, GA, Workshop D: Methodology for Comparing Risk Assessment, American Institute of Chemical Engineers, New York, NY, 1997, p. 16.

USE OF KEYPHRASE EXTRACTION SOFTWARE FOR CREATION OF AN AEC/FM THESAURUS

SUBMITTED: November 1999
REVISED: February 2000
PUBLISHED: March 2000 at http://itcon.org/2000/2/
EDITOR: B-C. Bjoerk

Branka Kosovac, PhD candidate,
University of British Columbia;
email:branka@civil.ubc.ca

Dana J. Vanier, PhD,
National Research Council Canada;
email:Dana.Vanier@nrc.ca

Thomas M. Froese, PhD,
University of British Columbia;
email:tfroese@civil.ubc.ca

SUMMARY: The paper describes a method used to collect terms needed for the development of a thesaurus in the roofing domain. This work is part of a larger effort to investigate the potential of thesauri as an aid in product modeling and as a tool for information management in model-based systems. Extractor, a software module that extracts keyphrases from documents, was used for collecting candidate thesaurus terms from Internet sources. The principal advantage of the Internet as a source of candidate terms is that it reflects the language that is actually used in communications concerning buildings and that it covers the widest range of different views on the domain. The advantage of using Extractor or similar software is that it allows processing huge text corpora available on the Internet while eliminating irrelevant terms. The methodology used was found to be highly useful, although it was not sufficient by itself for constructing a thesaurus for the architecture, engineering, construction and facilities management industries, as considerable human intervention was required. Some possibilities for customizing the software and for partially automating a thesaurus construction process are suggested.

KEYWORDS: thesauri, Internet, automatic indexing software, thesaurus construction.

1. INTRODUCTION

A thesaurus is a set of terms that are used in a specific domain of knowledge, "formally organized so that a priori relationships between concepts are made explicit" (Aitchinson and Gilchrist, 1987). Originally intended for indexing and retrieving documents, thesauri have increasingly been seen as knowledge bases and used beyond the domain of librarianship (Kosovac, 1998). The overall objective of this research direction is to investigate the potential of thesauri to assist the development and use of product models in the architecture, engineering, construction, and facilities management (AEC/FM) industries. It is believed by some (Vanier, 1994) that thesauri can help solve many of the semantic problems hindering the creation of robust product models for the industry and the efficient use of model-based systems. The goal of this project was the development of a prototype thesaurus in the roofing domain, and more specifically, for low slope roofs. One of the purposes was to explore the use of the Internet as a source of candidate thesaurus terms. The paper describes this particular aspect of the project — the method used to collect terms that are used in the low-slope roofing domain.

Extractor (Extractor 2000) is a machine-learning-based software module, developed by the Interactive Information Group of the National Research Council Canada, that scans an electronic document and extracts keyphrases best describing the document's subject matter. In this project, the authors used Extractor 2.0 as a support tool for collecting and selecting terms to be included in the proposed thesaurus.

In the process described, Extractor was used for a specific task and under given circumstances. Time and resource constraints did not allow full exploitation of Extractor's capabilities. These constraints also precluded

testing corpora (document collections) large enough to provide statistically valid results. Therefore, the work described is of explorative nature and cannot be considered as a study that evaluates the performance of Extractor. However, the patterns noticed in the analysis of the results can point to possible use of the software for related purposes and suggest possibilities for further research and development.

1.1 Problem description and general approach

Pragmatic analyses of the domain, available thesauri, and intended applications of the proposed thesaurus indicated that it should follow the format of the Canadian Thesaurus of Construction Science and Technology (TC/CS), a thorough, comprehensive, yet "dated" construction thesaurus (TC/CS 1978). One approach for developing a compatible microthesaurus is as follows:

- extract selected terms from the TC/CS to form a sectorial thesaurus,
- update the sectorial thesaurus to reflect current technology and terminology, and
- develop a micro-hierarchy of narrower concepts, if required.

Because of the "datedness" of the TC/CS (1978), it was decided that the following "bottom up" approach was more suitable:

- collect terms relevant to the field,
- select terms to be included in the thesaurus according to its intended purpose,
- check the terms against the existing TC/CS thesaurus,
- organize the terms into hierarchies following the TC/CS guidelines.

Therefore, the first phase of our work involved collecting candidate thesaurus terms. Typical sources for the initial collection of terms include:

- terminological sources in standardized form: existing thesauri, dictionaries, glossaries, classification schedules, encyclopaedias, lexicons, journal indices, back-of-the-book indices, term lists, treatises on terminology of a subject field;
- literature scanning;
- question (i.e., user query) scanning;
- users', subject experts', and compilers' knowledge (Aitchinson and Gilchrist, 1987).

Literature and question scanning play a crucial role for the usability of a thesaurus while the other sources serve mostly to clarify the meanings of terms, facilitate their arrangement into hierarchies, and fill gaps.

1.2 Use of information technology in thesaurus construction

Along with their use for thesaurus-management, computers have long been used for collecting thesaurus terms (Gilchrist 1971, Lancaster 1986, Aitchinson and Gilchrist, 1987). Their main use has been for automatic scanning of either titles and abstracts or full text documents from large textual databases and ranking the derived terms by frequency of occurrence. The explosive growth of the Internet has provided huge and diverse machine-readable corpora that have been used for a variety of purposes in computational linguistics and natural language processing. However, little work has been reported to date where Internet documents were used for extracting candidate thesaurus terms.

Since the mid-eighties, there have been considerable research efforts related to automatic extraction of semantic relationships and automatic construction of structured thesauri, e.g. Chodorow et al. (1985), Fox et al (1988). Despite some claims of successful R&D projects, such as MindNet (Richardson et al., 1998), wide use of full automation in thesaurus construction is still unrealistic. One simple but important reason is that the procedure requires special corpora — in almost all projects machine-readable dictionaries have been used. As a comprehensive, authoritative and up-to-date source for a certain subject domain is rarely available, it can be expected that computers will still be used only as support tools in this part of the process for a time to come.

1.3 Use of thesauri in product model-based systems

While there are many uses for technical thesauri, this project is aimed at exploring their application to product models and model-based integrated systems for AEC/FM: systems that communicate and work with object-oriented, semantically rich representations of project information. This paper focuses on the creation rather than the use of thesauri, but a few observations about the potential role for thesauri in model-based systems can be made.

Thesauri address the general task of mapping related concepts, a task that arises frequently in model-based systems. They provide formalisms and approaches for mapping concepts, such as the addition of relative specificity of relationships (e.g., one term might be described as related to, but more narrowly defined than, another term). Thesauri can be used to map not only related words and phrases, but to map a wider range of representational elements such as semantic models, heterogeneous forms of computer-based data, etc. Roles in which thesauri may be useful in model-based systems include the following:

- Basic terminology: Data models must use specific, unambiguous terms for each concept represented in the model. Real world language, and even the terminology used within the set of all AEC/FM computer applications, is much less precise, and a wide variety of terms may commonly be used to refer to the same concept. Thesauri can help map the terminology used by different applications and users to the terminology used within a shared product model.
- Schema: Information sharing in distributed model-based systems relies on common, standard data models (schemas). In practice, however, many different schema exist that must be mapped to each other. Examples include the mapping of application schemas to common shared schemas, accommodation of different views and perspectives representing various disciplines or users, and the generation or interpretation of new semantics from a project database. While research and development is progressing in the area of mechanisms for mapping schema (such as the Express-X language), several aspects of this mapping relate to thesauri issues and techniques.
- Heterogeneous representations: Within distributed model-based AEC/FM systems, information about a single physical component of a building may be represented in numerous forms, such as textual and numerical properties of a software object, data values dynamically calculated on-demand, 2D and 3D geometry in one or more formats (e.g., surface models, solid models, etc.), digital multi-media information such as photographs, hyperlinks to external references, and so on. Techniques must be developed to inter-relate all of these data forms at varying levels of granularity and to manage the heterogeneous data sets. Again, while this is beyond the scope of traditional thesauri applications, the underlying requirements are very similar and thesauri-based approaches are likely to provide useful contributions to the required capabilities.

Further work in this research direction, including the use of the developed thesaurus, will be reported in subsequent papers.

2. PROJECT

2.1 Approach

The (TC/CS 1978) is a huge thesaurus (approximately 15,000 terms) covering a wide subject field. Searching it for all terms relevant to the subdomain would be an expert-labour-intensive process of following links throughout different hierarchies and numerous general terms, and deciding which terms are relevant and current. Possibilities for automating this task are minimal. Another problem is that the TC/CS is rather outdated, especially considering the significant changes in the field of low slope roofing that occurred during the 1980s, with the introduction of new materials and types of roofing systems.

On the other hand, the TC/CS has a thoroughly elaborated structure and well defined inter-term relationships that facilitate the addition of new terms, given that their exact meaning is known. Furthermore, it is available in electronic form on the World Wide Web (http://www.cisti.nrc.ca/irc/thesaurus/.) thus allowing easy searching for known terms. For these reasons the "bottom up" approach suggested earlier seemed to be the most appropriate approach for this work.

As one of the primary intended uses of the proposed thesaurus is indexing and retrieval of Internet/intranet sources, the most useful source of terms would be corpora available on the Internet. The Extractor 2.0 documentation pointed to the suitability of the software for performing "literature scanning" of Internet sources as it integrates HTML and e-mail filters and permits processing of large corpora by extracting only relevant terms.

2.2 Preparatory tests

Since Version 2.0 of Extractor had been recently released at the time of this work (Version 5.1 has now been released), and since there were no similar previous efforts to build upon, the development of the methodology required some initial tests to roughly assess available sources and Extractor's behaviour. The tests were done on documents retrieved by general search services, such as Altavista (http://www.altavista.com) and the services listed below, and on documents from selected Web sites. The documents were processed by Extractor 2.0 varying the setting for the number of keyphrases to be extracted. After analysing Extractor's output, the results and the limited resources necessitated some modifications to the initially considered strategy.

First, the query initially used to test harvesting corpora using automatically generated queries that combine synonyms of the top term (*summum genus*) and its immediate narrower terms was as follows:

("low-slope roof" OR "flat roof*") AND ("built up" OR BUR OR "multi ply" OR "single ply")*

(where BUR is an abbreviation for built-up roofs). This query did not provide optimum recall and precision within some search services as processing a sample large enough to compensate for the deficiencies was unrealistic. The query was thus modified to the following:

("flat roof" OR "low slope") AND ("built up" OR BUR OR roofing OR membrane*)*

This better reflected the content and the language of the relevant documents. Although the new query did not eliminate all the "noise", nor did it ensure absolute recall, the retrieved sets of documents seemed acceptable for the purpose.

Second, the initial strategy was to process documents using the lowest setting for the number of keyphrases, identify terms that should be added as stop phrases (i.e., terms such as "roofs" or "roofing" that appeared as keyphrases in most of the documents preventing extraction of more specific terms), cluster documents based on the rest of the keyphrases, process them with the highest setting for the number of keyphrases, and repeat the process until the desired level of specificity was achieved. However, the inability to frequently customize the software by adding new stop phrases and the labourious task of processing the same documents more than once made this strategy unfeasible. Using the new release with the maximum setting for the number of keyphrases derived and simply removing the most frequent terms proved to yield satisfactory results.

It was observed that long documents, which tended to abound with very specific terms, yielded only very general terms in the Extractor output. Although keyphrases derived by Extractor reflected well the subject of the documents, they did not include specific terms that would be more useful for the purpose. Efforts to automatically divide long texts into meaningful sub-documents did not prove feasible with most of the Web documents. It was, therefore, done only on a small number of scientific papers that tended to be well organized and have a better HTML structure, thus allowing easy division by searching for heading tags.

2.3 Methodology

The immediate goal of the work is to extract relevant terms from Internet documents—not to evaluate Extractor 2.0. However, Extractor's performance has been continuously evaluated after each step in order to re-design the methodology or even give up the use of Extractor 2.0 if it would not produce the desired output. Though many of the tasks had to be performed manually, the methodology was structured in a fashion that would lend itself to automating the process or at least the use of clerical instead of expert labour.

2.3.1 Corpora

The following collections were selected to serve as the initial document set:

- Documents retrieved by general search services

The first 15 documents retrieved by advanced search with each of five major general search services (AltaVista, Excite, HotBot, Infoseek, and Lycos) were used, 75 documents altogether. Although most of the search services perform better with other search options, for consistency, the following Boolean query or the closest allowable option was used in each search:

("flat roof" OR low-slope) AND (built-up OR BUR OR roofing OR membrane*)*

The services were searched in alphabetical order, taking care to avoid duplicates. Where relevant documents from a certain site were grouped together, only the first one was used in order to avoid language of one author and frequent appearance of the same corporate names and trademarks.

- IRC Roofing Resources collection (Roofing Resources 1998)

This collection was included as it represents the core of the collection to be searched by the future thesaurus. Files bigger than 20 KB (arbitrarily established limit) were divided by headings to form 40 documents for the extraction of keyphrases.

- Relevant documents retrieved at FacilitiesNet (FacilitiesNet 1998)

The criterion for the inclusion of this site was its high content of documents relevant to the facilities-management aspect of flat roofs. This aspect is represented only in a minor portion of documents available on the World Wide Web but is important to the wider context of this work. Sixty-one relevant documents were retrieved and processed.

- Collection of selected articles

This collection was compiled by following links from various lists of relevant sources but without aspirations to be comprehensive, exhaustive, nor of highest quality. It has been included to allow comparison of the results with those achieved by automatic harvesting of sources using general search services.

Messages from newsgroups archives were not processed separately but a small number of this type of document was included in AltaVista hits.

2.3.2 Procedure

Each document was processed by Extractor 2.0 with the number of keyphrases set to maximum. . The extracted terms were gathered in a list. The following information was recorded for each term in the Extractor list:

- relevance factor number (provided by Extractor),
- document from which the term was extracted,

and for each document:

- collection ID,
- size of the file.

After processing each set of documents, the results were analyzed, compared with other sets, and the sets were integrated and processed together. The final set of keyphrases was reviewed and compared to the (TC/CS 1978) and existing glossaries.

2.3.3 Criteria

Single- and multiple-word terms were not treated separately, as the list was of modest length. Singular and plural forms of the same term were counted together but terms that may have different meanings when used in plural (e.g. roofing--roofings) were specially marked [PL]. The terms were first searched in TC/CS for exact matches [=]. Qualifiers from the thesaurus (i.e. auxiliary terms used to make the meaning of the term unambiguous) were ignored in this step. Terms identified as general terms in the thesaurus, i.e. terms that do not have a fixed hierarchical level in TC/CS but can be associated with terms of varying degrees of specificity (e.g. "life" or "requirements") were additionally marked [GT]. So were those that had a further developed hierarchy of narrower and/or part terms in TC/CS [+].

After the identification of exact matches, terms from the list were also searched for occurrence as:

- [PH] phrases from the thesaurus
- [Q] qualifiers in the thesaurus.
- [~] close matches, meaning terms having the same stem.

Analysis of the remaining terms identified:

- [A] acronyms;
- [N] proper names;
- [*] phrases ill defined by Extractor (meaning that they could not stand alone either as terms in the thesaurus—which includes only nouns and noun phrases—or as meaningful keywords for a document—e.g. "requiring", "single", or "install"); and
- [$] matches that have the same form but different meaning in the thesaurus.

The appendix A contains a marked list of terms that were extracted as keyphrases from more than two documents.

3. RESULTS AND DISCUSSION

Terms with extremely high frequency of occurrence, which should have been made stop-phrases, were identified early in the process. These terms were the same for each set of documents and also could have been identified by processing the list itself in Extractor 2.0 with the number of keyphrases set to the minimum..

The results from the general search services were then compared with those from selected collections. There were no significant differences noted in the relevance of extracted keywords that would justify the laborious task of searching, evaluating, and selecting sources. The quantity and diversity of documents that can be easily retrieved by general search services can successfully compensate for the quality of selection. As the first 20 documents from the compiled collection described in section 2.3.1 did not bring new terms to the list, this collection was not further processed and it is not included in the final results. However, the documents have been saved for later comparison of scholarly and natural language terms and for exploration of specific sub-areas.

The final list consists of 1054 terms (2423 occurrences) extracted from 176 documents. Almost half of the terms extracted were single words (49 %). They accounted for almost all of the top 4% most frequent terms with only two exceptions that would normally be included in stop-phrases. The usual practice of treating such terms separately was not followed as most of the terms were also identified as single word terms in the TC/CS.

A huge number of single occurrences of a term (78%) can be explained by the insufficient size of the sample. In order to extract relevant terms from this group, they were searched for component words and stems that could also be found in other terms. Terms found in this way were ranked higher in the list as more relevant. Rough scanning of the remaining single-occurrence terms found very few terms that were relevant to the field. For that reason, these terms were excluded from further processing. It is important to note, however, that this group contained a substantially smaller percentage of terms marked as ill-defined, significantly contributing to the overall average of only 1% ill-defined terms.

3.1 Comparison with the existing thesaurus

Checking the final list of terms against the TC/CS revealed a high relevance of the terms extracted by Extractor 2.0 to the field. Among 130 terms that appeared more than twice, 56% had exact matches in the thesaurus, 12% were found only in phrases, and 6% were marked as close matches. This breakdown is illustrated in Fig. 1. Terms that occurred twice had even more exact matches (63%) in the thesaurus but were less frequently found in phrases as they included more multi-word terms.

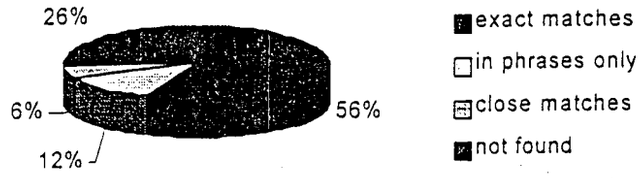


FIG.1: List of terms with more than two occurrences compared with TC/CS

Fig. 2 shows the breakdown of terms with more than two occurrences that had no matches or close-matches found in the existing thesaurus — 12% (or 3% of all the terms extracted) were proper names, 8% (2% of all the terms) were acronyms, and 19% (or 5% of all the terms) were marked as ill defined.

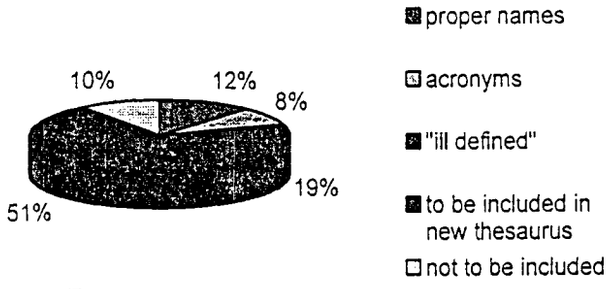


FIG. 2: Breakdown of terms not found in TC/CS

The majority of mismatches, however, do not indicate irrelevance of the terms but more often the outdatedness of the TC/CS. The frequent occurrence of the term "membranes" for example, and the phrases containing "membranes" that are not found in TC/CS reflects changes in the field of low slope roofing and its terminology. The noise-making terms, i.e. extracted terms that do not belong to the low-slope roofing domain, come from specific kinds of documents, mostly glossaries and book catalogues. Such documents can be easily excluded from the beginning by modifying the initial query.

The relatively high coincidence of terms may, on the other hand, indicate a lack of more specific terms that would be required for developing a microthesaurus. Whether this is the case can be established only later in the process.

The matches were found in all semantic classes and in various hierarchies, showing a broad coverage of the domain. Completeness of coverage, however, is yet another problem that cannot be properly evaluated at this point but only after organizing terms into hierarchies (Petersen 1990).

3.2 Comparison with the glossary

The terms were also matched against relevant (i.e. low-slope-roofs related) terms extracted from a roofing glossary (Biegel 1989). The coincidence was significantly lower; only 31 % of the glossary terms were found in the Extractor's output. The unmatched terms were mostly very specific terms (e.g. "back-nailing" or "cutoff"). Since such narrow concepts usually do not represent a document topic, it cannot be expected that keyphrase generation software will retrieve the terms denoting them.

4. CONCLUSIONS AND RECOMMENDATIONS

4.1 Observations on performance of Extractor

Extractor 2.0 appears to be a suitable tool for collecting thesaurus terms from the Internet. It can be used on any PC running Windows 95/98/NT, or on Linux or Solaris systems. Although its use in this project required extensive manual work, it is estimated to be more efficient than both manual and automatic full-text literature scanning. The principal advantage is that it allows scanning and handling of a significantly larger number of documents, thus providing better coverage of the field and its terminology. In addition, Extractor can be easily embedded in other software. This feature, which was not used in this project, allows further automation of many of the described tasks.

The number of phrases marked as ill-defined (see the explanation in section 2.3.3) was 1% of all the extracted terms. Since the total automation of the thesaurus constructing process is not considered and since ill-defined phrases cause no serious consequences, this percentage can be considered negligible. Therefore, increasing the number of keyphrases even above the maximum that was possible in version 2.0 (in order to retrieve more specific terms) would probably be safe. In most cases the lack of more specific terms in the Extractor output will not represent a deficiency; terms too specific to be the subject of a document are rarely included in a thesaurus. In addition, their presence might make one of the most important decisions in thesaurus constructing — where to stop — even more difficult. However, if constructing a microthesaurus, or if for any other reason more specific terms are needed, these terms may be obtained by processing larger corpora or by narrowing the searches for the Internet documents to be processed.

4.2 Observations on the Internet as a source of thesaurus terms

The Internet represents an extremely useful source of thesaurus terms. It provides huge corpora covering numerous aspects of a domain and different vocabularies—from the highly scholarly to the most informal. Internet documents reflect the language that is current, actually used in the field, and most likely to be used in queries. The results also showed that documents randomly harvested using general search services could provide equally valuable terms as controlled subject collections. The collected documents can be further analysed to complement the results of the described methodology.

4.3 Implications for the further work on the thesaurus

After establishing relationships between terms and organizing them into the hierarchies of the existing thesaurus, it can be expected that certain areas would need further development. Upon identification of such areas, gaps will first be filled with the following:

- terms derived by Extractor from new sets of documents retrieved by more specific terms,
- terms manually extracted from documents already retrieved and judged as relevant to the subfield according to keyphrases derived by Extractor.

The use of these two sources is prioritized for the reasons listed in section 4.2. However, these sources cannot ensure comprehensive coverage of the domain. Therefore, manual extraction of terms from alternative sources will probably be needed. Encyclopaedic and textbook type documents, roofing manuals, various kinds of term lists, and architectural details' labels are expected to best serve the purpose.

4.4 Possibilities for automating the process

Some of the tasks that could be fully or partly automated in applications used for similar purposes include the following:

- automatic retrieval of relevant documents from the Internet and their processing with Extractor,
- periodical processing of the list by Extractor for finding terms with significantly high occurrence and making them stop-phrases,
- ranking terms by frequency of occurrence,
- exclusion of terms that occur only once in large corpora,

- automatic exclusion of geographic names ,
- grouping of terms containing the same words or stems,
- grouping of terms by co-occurrence in documents,
- suggesting inter-term relationships by co-occurrence and syntax.

5. FINAL NOTES

The work described was carried out in February 1998. In the meantime, a 336-terms pilot thesaurus has been developed (http://www.nrc.ca/irc/thesaurus/roofing). As these terms represent only a very small portion of the domain and of the terms collected in this process, final conclusions on the usefulness of the methodology cannot be drawn yet.

Since the time of the study, several new versions of Extractor have been released. Anyone interested in implementing the method described in this paper should review new features of the software by visiting the Extractor Web site or by contacting the software's authors.

Original data and full results of the study are available from the authors of this paper.

6. ACKNOWLEDGEMENTS

The authors wish to thank the Institute for Information Technology of the NRCC for the use of Extractor 2.0, and more specifically, to thank the software's author, Dr. Peter Turney, for his support and encouragement. The authors also wish to acknowledge the Canadian Institute for Scientific and Technical Information and the Institute for Research in Construction, both of the NRCC, for their financial contributions to this research. The authors thank both Dr. Ferrers Clark and Mr. Scott Mellon from these respective organizations for their support and encouragement during the course of the research. The authors also acknowledge the work of Prof. Colin Davidson and the IF Group in the development of the TC/CS Thesaurus.

7. REFERENCES

Aitchinson, J. and Gilchrist, A. (1987). *Thesaurus Construction*, Aslib, London, UK.

Biegel, S. (1989). Roofing Materials, *Encyclopedia of Architecture, Design, Engineering & Construction*. Vol. 4, American Institute of Architects, 314-319.

Chodorow, M., Byrd R. and Heidorn G. (1985). Extracting semantic hierarchies from a large on-line dictionary, *Proceedings of the 23rd Annual Meeting of the ACL*, 299-304.

Extractor (2000). National Research Council of Canada, Interactive Information Group, Ottawa, Canada. Available from: http://extractor.iit.nrc.ca. [Accessed February 7, 2000]

FacilitiesNet (1998). Trade Press Publishing, Milwaukee, WI, USA. Available from: http://www.facilitiesnet.com. [Accessed November 27, 1999]

Fox, E.A., Nutter J. T., Ahlswede T., Evens M. and Markowitz J. (1988) . Building a large thesaurus for information retrieval, 2nd Conference on Applied Natural Language Processing, Association for Computational Linguistics, (Ballard B., ed.), Bell Communications Research, Morristown, NJ, 101 -108.

Gilchrist, A. (1971). The Thesaurus in Retrieval, Aslib, London.

Kosovac, B (1998). *Internet/Intranet and Thesauri*, Canadian Institute for Scientific and Technical Information, Internal Report, National Research Council Canada, Ottawa, Canada. Available from: http://www.nrc.ca/irc/thesaurus/roofing/report_b.html [Accessed November 27, 1999]

Lancaster, F.W. (1986). *Vocabulary Control for Information Retrieval*, Information Resources Press, Arlington, VA, USA.

Petersen, T. (1990). Developing a new thesaurus for art and architecture, *Library Trends*, Vol. 38, No. 4, 644-658.

Richardson, S. D., Dolan W. B. and Vanderwende L. (1998) . MindNet: acquiring and structuring semantic information from text. Microsoft Research Technical Publications (MSR-TR-98-23). Available from ftp://ftp.research.microsoft.com/pub/tr/tr-98-23.doc [Accessed January 31, 2000]

Roofing Resources (1998). National Research Council of Canada, Institute for Research in Construction, Ottawa, Canada. Available from: http://www.nrc.ca/irc/roofing/roofing.html. [Accessed November 27, 1999]

TC/CS (1978). *Canadian Thesaurus of Construction Science and Technology*, Department of Industry, Trade and Commerce, Government of Canada, Ottawa. Available from: http://www.cisti.nrc.ca/irc/thesaurus/. [Accessed November 27, 1999]

Vanier D.J. (1994). Canadian thesaurus of construction science and technology: A hypercard stack, *Proceedings of the Joint CIB Workshops on Computers and Information in Construction (Montreal, Que., Canada)*, (*CIB Proceedings*, Vol. 165), 559-564.

APPENDIX A

TABLE 1: List of terms extracted as keyphrases from more than two documents Sorted by frequency of occurrence indicated by the number in the first column.

	TERM	=	~	GT	+	PH	Q	A	N	*	PL	S	NOTE
158	roofing(s)	=									PL	S	
139	roof(s)	=											
74	membrane(s)					PH	Q						
50	materials	=											
32	installation	=				PH					PL		installation(activity)
32	products	=				PH							products(agents)
27	performance	=				PH							
27	water	=				PH							
26	design	=				PH					PL		
25	insulation					PH							
25	roofing system(s)												
24	requirements	=		GT		PH							
22	asphalt	=											
19	deck(s)					PH							
19	repair		~										repairing
17	maintenance	=	~		+								maintenance(restoring)
17	manufacturer	=											
16	flashing(s)	=			+								
14	coating(s)	=									PL		coating(process)
13	fasteners					PH	Q						
13	inspection	=		GT									
13	joint(s)	=			+								joints(junctions)
13	specifications	=		GT		PH							
12	BUR							A					built up roofings found
12	costs	=											
11	construction												
11	felts	=				PH						S	
11	structures	=											structures(buildings) structures(construction) structures(non building)
11	waterproofing					PH							
11	wind	=				PH							
10	flat roof(s)	=											
10	properties	=		GT							PL		property(quality)
9	components	=					Q						
9	moisture					PH							
9	sheets	=			+								sheets(shape)
9	slope(s)	=					Q						
9	standards	=				PH							
8	industry	=				PH							
8	install									*			
8	projects			GT		PH							
8	replacement					PH							replacement value
8	resistance	=				PH							non-descriptor
8	walls	=				PH							
8	warranties	=				PH							

7	air	=											
7	ballast	=										S	ballast(gravel)
7	building owners	=											
7	flexible membrane												
7	market(s)	=											
7	modified bitumen(s)												
7	panel(s)	=											
7	seams					PH							
6	applications	=				PH						S	
6	built-up roofing	=											
6	consideration												
6	drain(s)					PH							
6	evaluation	=		GT									
6	installing									*			
6	life	=		GT		PH							
6	load(s)	=				PH							
6	protected membranes												
6	roofing contractor(s)												contractors
6	shingles	=											out of scope
6	substrate												
6	temperature	=		GT									
6	vapour barrier(s)	=			+								
5	assemblies	=											assemblies(components)
5	base flashing(s)	=										S	
5	Bitumen(s)	=											
5	built-up roof		~										built up roofings
5	condition(s)			GT							PL		
5	drainage	=				PH							
5	EPDM		~										epdm rubber
5	experience	=										S	ND for "skill"
5	facilities					PH	Q						
5	leak		~										leakage
5	low slope												
5	pressure					PH							
5	preventative maintenance												but *prevention & maintenance*
5	protection	=		GT									
5	roof decks		~										roof deckings
5	single									*			
5	thermosets		~										thermosetting resin (ND)
4	built-up									*			
4	coal(-)tar	=											
4	damage	=											
4	flow	=											flow(fluids)
4	layer					PH						S	
4	movement					PH							ND for *motion*
4	recommendations	=		GT									
4	reinforcing		~			PH						S	
4	requiring		~									S	
4	Roofing Contractors Association								N				
4	roofing membranes												
4	thermal insulation	=											
4	vapour retarders												
3	Aduron								N				
3	architect(s)	=											
3	barrier					PH						S	ND
3	base coat(s)											S	
3	building envelope												
3	Canada								N				
3	coating system												but "coating(process)"
3	condensation	=				PH							
3	contract(s)	=											
3	control	=		GT									
3	counter(-)flashing(s)												
3	cracks	=											cracks(fissures) ND
3	differential movement(s)	=											ND
3	edge(s)	=		GT		PH							edge(boundary)
3	equipment	=											
3	flat					PH				*			

3	foot					PH							
3	gravel	=											
3	installer					PH							
3	liner												
3	measurement(s)	=		GT		PH							
3	metal roof		~										metal roofings
3	minimum					PH							
3	National Research Council								N				
3	owner(s)	=				PH							
3	parapet					PH							
3	polymer	=											ND for "polymeric materials"
3	reinforcement					PH							
3	review												irrelevant
3	SPF							A					
3	steel	=											
3	surfacing	=											
3	waterproofing membrane		~										waterproof membranes (ND)

PL terms may have a different meaning when used in plural
= exact match found in TC/CS
~ close match found in TC/CS
GT term represents a general term in TC/CS, i.e. term that does not have a fixed hierarchical level but can be associated with terms of varying degrees of specificity
+ term has further developed hierarchy of narrower and part terms in TC/CS
PH word found as a part of phrases included as terms in TC/CS
Q term used as a qualifier in TC/CS
A acronym
N proper name
* term "ill-defined", i.e. cannot stand alone as a thesaurus term
S matches that have the same form but different meaning
ND non-descriptor, i.e. term pointing to a descriptor (authorized term) used as a thesaurus entry

Deriving Concepts Hierarchy

Yousef ABUZIR
AlQuds Open University
Jerusalem,Palestine
yabuzir@qou.edu

Abstract. Information Retrieval (IR) covers the problems relating to the effective storage, access, searching and locating documents that are relevant for user's information need or query from large collection documents. Many techniques and tools have been developed to improve these processes. One of these tools is the thesaurus. This paper will present a tool for users to build thesauri according to their own requirements. The objective of this paper is to present an overview of how thesauri can be created automatically. The main goal is to design a comprehensive thesaurus-building tool that can be used in any specialty field. The system provides sufficient structure to represent different relationships among concepts and terms as well as to extract these relationships. In addition a thesaurus builder can also use different methods, to extract terms and relation between them from document corpus. So, the system provides a set of interface functions through which the user can easily build a thesaurus.

1 INTRODUCTION

A thesaurus is a set of concepts in which each concept is represented with at least synonymous terms, broader concepts, narrower concepts, and related concepts. A term is a word or sequence of words that refers to a concept [1]. The relations or links broader, narrower, related, and synonymous have been defined for thesauri [2, 3]. Originally intended for indexing and retrieving documents, thesauri have increasingly been seen as knowledge bases and used beyond the domain of librarianship.

Thesaurus has long been a concern in lexicography, and recently, it has found many applications in machine translation, information retrieving, and computational lexical semantics, etc. [4,5,6]. A thesaurus is a structured system of terms encoding explicitly semantic relations like synonyms, hyponyms, hyperonyms, part-whole relations, associations, etc. Because manual construction, maintenance, and updating is highly time-consuming, in the recent years methods have been developed that automatically construct a thesaurus out of a collection of documents related to a given subject. These methods are based on statistical or linguistic treatment of the document collections. Within this context the task of extracting terminological terms out of a text arises. Thesauri constructed in this way are in general less structured.

A thesaurus is a valuable tool in Information Retrieval (IR), both in the indexing process and in the searching process, used as a controlled vocabulary and as a means for expanding or altering queries. Most thesauri that users encounter are manually constructed by domain experts and/or experts at document description. Manual thesaurus construction is a time-consuming and quite expensive process, and the results are bound to be more or less subjective since the person creating the thesaurus make choices that affect the structure of the thesaurus. There is a need for methods of automatically construct thesauri, which besides from the improvements in time and cost aspects can result in more objective thesauri that are easier to update. Therefore, building thesauri automatically from corpus received a large attention in recent years.

In the second section of this paper, we summarize the research issues involved in the general problem examined in this paper, as well as the solutions proposed for several of these issues. The problem of thesaurus construction can be broken down to the following sub-problems term extraction and term organization - deriving relationships between terms from texts. Section 3 presents a general description of the Thesaurus Construction Tools and the techniques used. Sections 4 and 5 present experimental results of the different models for term extraction and building the hierarchical relationships

between terms. Section 6 discusses the conclusions and future work.

2 THESAURUS CONSTRUCTION

The most conventional, approach to thesaurus construction is to build it manually. This labor intensive and hence only possible in specialized domains where repeated use may justify the cost. This approach requires a domain expert to prepare a hierarchical structure of topics relevant to a particular subject area. Searchers then employ terms from this hierarchy to form queries that automatically expanded to more specific or general terms.

The growing amount of information and the need for quick access to it slowly changes the significance of Information Retrieval (IR) systems. In order to turn Information Retrieval systems into more useful tools for both the professional and general user, one usually tends to enrich them with more intelligence by integrating information structure, such as thesauri. Since it is difficult and expensive to build thesauri manually, many researchers attempted to construct thesauri automatically. This section, discusses the different approaches to thesaurus construction. It also describes some systematic procedures for thesaurus construction.

The construction of thesauri generally takes the following form:
Extract word co-occurrences.
Define similarities (distances) between words on the basis of cooccurrences.
Cluster words on the basis of similarities.

2.1 The Different Approaches

There are different approaches to construct thesaurus automatically. Figure 1 shows these different approaches. We can classify these approaches into two groups. In the first one we can group the different approaches according to the source of terms used to construct the thesaurus. The later is based on the corpus-based techniques.

2.1.1 Approaches Based on Source of Data

There are three approaches to construct a thesaurus based on the soured of terms. The first approach, on designing a thesaurus from document collection, is a standard one [7, 8]. Here the idea is to use a collection of documents as the source for thesaurus construction. This assumes that a representative body of text is available. By applying **statistical** or **linguistic** procedures we can identify important terms as well as their significant relationships.

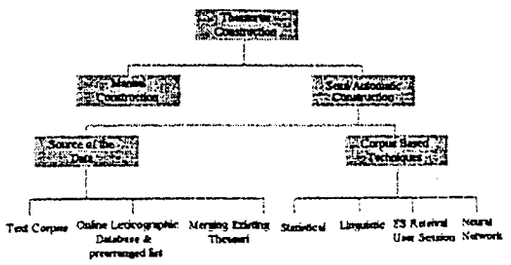


Figure 1 The Different Approaches to Thesaurus Construction.

The simplest approach is to reuse existing online Lexicographic database, such as wordNet or Longman's subject codes [9].

Pre-arranged list of terms like table of content, index of book or other format can be used to construct thesaurus automatically [10].

The third approach is merging existing thesauri. This third approach is appropriate when two or more thesauri for a given subject exist, that need to be merged into a single unit [11, 12].

2.1.2 Approaches Based on Collection Treatment

In the second group, the classification based on the different corpus techniques used to build thesauri automatically. A corpus-based method performs a computation on the text of the documents in the corpus to induce a thesaurus.

Subsumption - construct a hierarchical thesaurus from a computed list of complex noun phrases where subsumption roughly corresponds to the subset relation defined on terms (i.e.; "intelligence" subsumes "artificial intelligence") [13].

Linguistic - Several researchers have used head-modifier relationships to determine semantic closeness [14, 15]. Co-occurrence statistics and head-modifier relationships get at different kinds of information. Two terms that modify the same words (or are modified by the

same words) often belong to the same semantic category.

Statistical - approaches based on semantic relatedness [16] by considering the occurrence of terms in document. Documents are clustered into small groups based on a similarity measure that considers two documents similar if they share a significant number of terms, with medium frequency terms preferentially weighted. Terms are then grouped by their occurrence in these document clusters. Other statistical approaches construct thesauri by transposing the standard term-by-document matrix in order to define a similarity measure on terms rather than documents [7, 17].

Expert System -The third automatic is based on tools from expert systems [18]. In this approach thesauri are built using information obtained from users. For example, if a retrieval system user combines two terms by OR, in his/her query these terms are probably synonyms.

Neural Network - The forth approach is based on Self-Organizing Map (SOM). SOM is an unsupervised two-layer neural network used to summarize large high-dimensional data. Roussinov and Chen explained how they used Kohonen's SOM as a tool for extracting semantic relationships between words and creating a hierarchy of categories [19]. Their prototype system creates a hierarchy of concepts from document returned by a WWW search engine using SOM.

3 AN OVERVIEW OF THE SYSEM

This section will present our system, which is a thesaurus construction system that can be used as a tool to build domain independent thesaurus.

3.1 General

Our system is a tool for users to build thesauri according to their own requirements. The main goal is to design a comprehensive thesaurus-building tool that can be used in any speciality field. It provides sufficient structure to represent different relationships among concepts and terms as well as to extract these relationships. In addition a thesaurus builder can also use different methods to extract terms and relation between them from document corpus. So, the system provides a set of interface functions through which the user can easily build a thesaurus.

The system provides two different modes depending on the user. These two modes are training and expert mode. In the first, users can learn and experiment the different methods for automatic term extraction and finding relationship between these terms. The second one can be used or assisted expert who is interesting in building domain independent thesaurus.

The system allows easy to extract the required terms. It also provides the ideal environment to build the hierarchical relations between these terms. It contains different Models that can be used or assisted in thesaurus construction (see Figure 2):

Term Selection Models – the extraction of terms (indexing terms) from text document. Automatic term extraction approaches like Document frequency (DF), Inverse like Document frequency (IDF), Discrimination Value (DV) and Frequency of Occurrence are used to find index terms. These statistical models predict the strengths of terms from the frequencies of these terms in the collection.

Terms co-occurrence Models - the occurrence of two terms or more can be perceived as an evidence of a relation between these words. The text documents are read in word by word. Whenever one of the keywords occurred with the word in query, the sentences are displayed and statistical data are stored to find the strength of the relation between terms. Term co-occurrence used to produce structures or maps of related terms.

Hierarchical organization of terms Models – Models for building hierarchical relationships between terms from text. Three models support creation of hierarchical relation are used, the first model that produce concept structure with an ordering from general terms to more specific that use cohesion statistic to measure the degree of association between terms. The second one is use the lexico-syntactic Analyzer to find the relationship between terms. The other one is use the subsumption.

HTML Hierarchical Model - to build a hierarchical structure of terms automatically from a set of HTML documents.

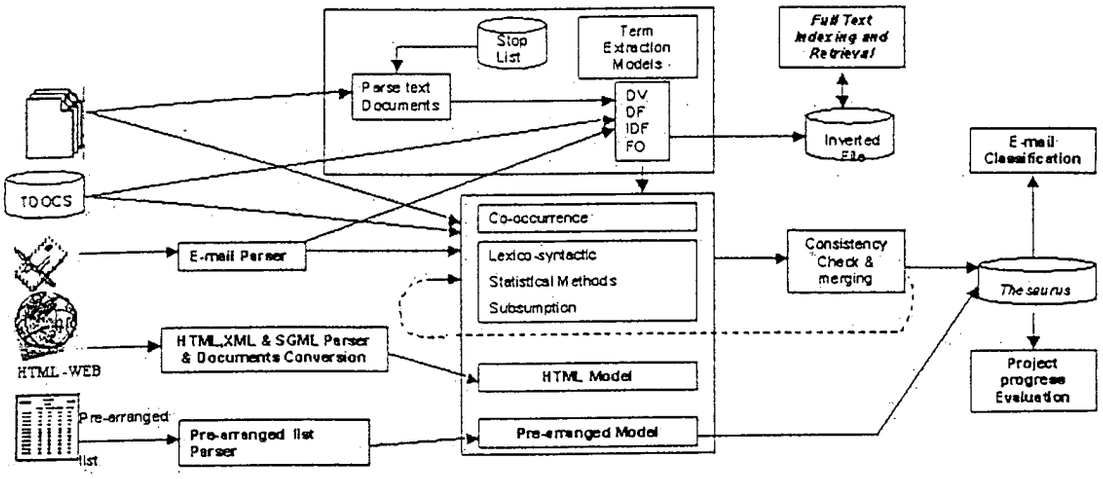


Figure - 2 Architecture of the System

ThesConv Model – Thesaurus Converter has import function to convert different input formats like book's index to TDOCS [20] Thesaurus Manager Data Base and spreadsheet. The result database is powerful for processing thesaurus relations.

Full search and retrieval Model - This model retrieves the required documents from the collection. A list of documents names can be obtained by giving search term(s). User can browse the selected documents.

Documents Conversion Models – This Model converts HTML, SGM and XML document to ASCII format. Other application can extract email messages and does information extraction from the email messages. Sender, subject, etc, are extracted from the email message.

3.2 Term Extraction

The objective of this tool was to automatically build domain independent thesaurus from a set of documents. . This translates into three basic phases: term extraction, relationships deriving and filtering and reviewing of the terms and relationships.

The problem of thesaurus construction can be broken down into the following sub-problems:

- Term extraction.
- Term organization - deriving relationships between terms from texts.

Term extraction: terms for the thesaurus were to be extracted from the documents and had to best reflect the topics covered within them.

Relationships deriving: these models provide the user with the relationships between terms

The final stage is the filtering and reviewing of these relationships. Before converting these results into the database of the thesaurus the user had to remove the noisy relationships that are in the result.

Some terms extracted from document text by Term Extraction Model may not function effectively as indexing terms in a thesaurus. A decision must be made about what terms should be included in the thesaurus. To decide what to include, various methods like Document Frequency (DF), Inverse Document Frequency (IDF), Discrimination Value (DV), and Frequency of Occurrence are used.

4 Lexico-syntactic Model

In order to extract terms and relations between terms, different methods can be used. One of these methods requires predefinition of lexico-syntactic patterns as well as manual traverses on outputs of terminologists, in order to find pairs that belong to the predefined relations. Hearst [21, 22] reports a method using lexico-syntactic patterns to extract lexical relations.

The patterns were constructed during manual analysis of text documents in the area of Total Quality Management (TQM). A sample of 48 different documents relates to Total Quality Management was used. Some patterns may simply not be applicable to a particular Broader-

Narrow Term (BT-NT), while in other cases there may be a natural default pattern.

Table 1 - Examples of the Patterns		
Pattern	Pair of term (BT-NT) Relationships	A Sample Example
for	(system, organization) (system, Management system) (process, Project) (document, record)	It is a system for translating the strategic intent on an organization to the managerial and operating spheres of decision making.
As	(system, organization) (system, process) (supplier, service)	He views the organization as a system and advocates using a scientific method to optimize the system

The experiment was meant to find positive evidence of various patterns. The system uses one pair of terms at a time to extract sentences that contain this pair. For each pair a list of sentences where the pair occurred in the text was made. From these lists common patterns were identified. In some cases a particular patterns included more that concepts (pair of terms), as can see in Table 1. Table 1 includes some examples of the patterns. The pattern formats extracted in our experiment are primitive. An example, in the first case we tried (system, organization) pair, and with just this pair we found new patterns. Next we tried other pair and discovered more patterns.

By using the Document Viewer of the system all keywords terms are directly visible. The system combines ideas of term identification, relationship extraction for terms and label these terms and relationship. The system requires only few manual definitions of lexico-systactic pattern and the need to know these lexico-systactic patterns in advance.

At the third step, a set of sentences is extracted. These sentences are lemmatized, and terms are identified. So, we represent a sentence by a lexico-syntactic expression. For example, the relation (*system, management system*) allows extracting the following sentence from the corpus TQM:

"However, in their study of target analysis system at Nissan, Balachandran and Srinidhi (1991) found the use of target analysis more as a management system to manage not only costs but also quality."

From this sentence, we produce the lexico-syntactic expression: Term1 as Term2 but also Term3.

A lexico-syntactic expression is composed of a set of elements, which can be lemmas, punctuation marks, or words, such as TERM1, etc. Through this simplification process, we have a more generic representation of relevant sentences, and comparing these sentences is easier.

The following sentences (Table 2) extracted from TQM corpus show other examples of the hyponym relation between terms:

Table 2 Examples from TQM documents	
Thus, suppliers include business such as distributors, dealers, warranty repair service, transportation, contractors, and franchises, as well as that provides materials and components.	Term1 **such as** Term2, {$Term_i$,}and $Term_n$
Supplier also include service suppliers, such as health care, training, and education providers.	Term1 *also include* Term2

The system supports interactive extraction of the terms and relations using the Lexico-Syntactic Model. With just the previous Lexico-syntactic patterns we find more new patterns as well as new terms (Figure 3). Note that for these patterns, even though they have emphatic term, this does not affect the fact that the relation indicated is hyponymic. The application of this technique to TQM domain showed great success.

5 Deriving Concept hierarchies from HTML

5.1 Introduction

All the approaches to thesaurus construction mentioned so far perform what we might call

construction by content, since information for constructing a thesaurus is extracted from the text of the document. Thesaurus construction by content does not exploit an essential aspect of a hypertext environment like the Web, namely the structure of documents and the link topology. In this section we investigate a technique for automatic thesaurus construction, which we have called construction by context, since it exploits the context in the HTML document to extract useful information for building thesaurus. Thesaurus construction by context exploits relevance hints that are present in the structure and topology of the HTML documents published on the Web. Combining a large number of such hints, an adequate degree of accuracy can be achieved in constructing the thesaurus.

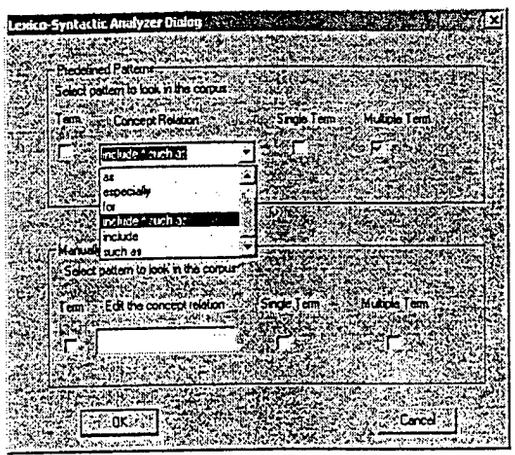


Figure 3. Lexico-Syntactic Dialog

The focus of this section is the construction and evaluation of an automatic tool for thesaurus construction from HTML pages. The system is a tool which, given a specific URL address of the web page that is broad and well-represented of the domain, will find out and return a list of terms that it considers the most authoritative for that topic. It is built to perform a local analysis of text, meta-tags and links to arrive at a "global consensus" of the best terms and relationships between them for specific domain.

5.2 System Overview

Figure 4 shows an overview of the construction process, which is made up of several steps. First the application accepts URL address as input and the HTML page is fetch.

From this HTML document plain text, meta-data like meta-tags, titles, headings, list, etc and URL are extracted. The outcome is stored in a database and used in later stages.

For each HTML document, the tool parses the text and classifying it into a maximum of three different groups. The groups give a local analysis on text, met-data and links found in the HTML page. These groups are (see figure 5):

- Metadata: as present in HTML META-tags and important text, like document title and all HTML headings.
- Plain text: all other text.
- URL list: the list of the URL in the HTML page.

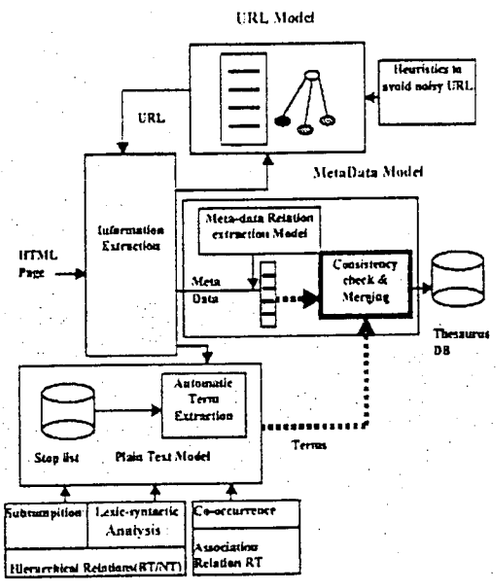


Figure 4 Architecture of the HTML Model

5.3 URL Extraction Model

Any URL found during the parsing is passed to processing if it points to a document within the current site, and stored for later analysis if it points to an external site, currently limited to depth of 3. This allows performing a depth-first visit of a site, collecting any categorization information it contains about itself and other sites.

Quite often the HTML pages of a site have a characteristic structure represented by links across pages. For instance there can be references to the main page, or links to the general services available in the site, like searching within the

site, help or information pages. Finally, there can be advertisement banners in precise positions in each page. We want to avoid processing such pages. In order to identify these structural links, we used some heuristics about the structure of URL and an initial analysis of pages reachable from the starting page. These links placed in a stop list of URLs and discarded in the subsequent analysis of the site.

This task starts from a list of URLs, retrieving the documents referred by each of them and analyzing the structure of the document expressed in terms of its HTML tags.

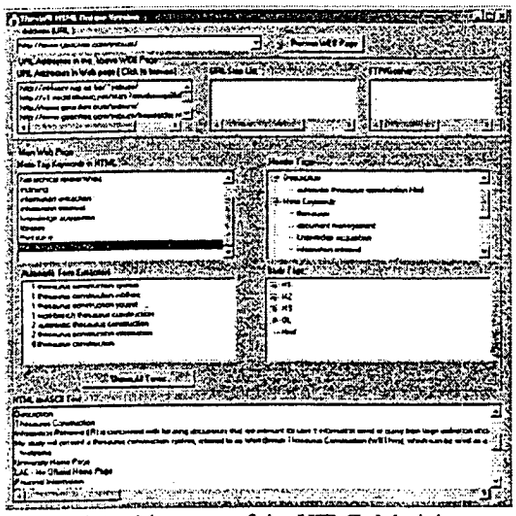


Figure 5 Architecture of the HTML Model

5.4 Meta-Tags Analysis

The tags considered are currently <TITLE>, <Hn>, <UL>, <OL>, <A>, etc. and a meta-tags <META Classification Keywords>. Whenever one of these tags is found, a context phrase is recorded, which consists of the title within a pair <Hn> </Hn>, or the first portion of text after a <UL> or <LI> tag, or the phrase within a <A> tag.

In the analysis, tags related to layout or emphasis (<EM>, <B> <I>, <CENTER>, <FONT> etc.) are discarded here. Such tags can be effectively used in plain text analysis to extract thesaurus terms.

In general the system will build the structure of the web pages as Follows:

1. Accept a URL, i.e. HTML document.
2. The system defines a target node using a start tag and an end tag, start tag is used to label each node.
3. Define a target low-level node within the node.
4. Repeat step 3 if more low-level nodes to be defined.
5. Repeat steps 2-4 if more nodes to be defined.

In order to extract the relationship between terms, the system exploits the hierarchical structure and sets of extraction rules. For each node in the structure, it is associated with rules that extract that particular information or element.

During the parsing process, the system applies text extraction rules for each type of tags. There are rules to extract and build the tree structure of the tags <Hn>. Such a tree has the tag <TITLE> as a root. The extraction rule for each of these tags is applied until all tags have been extracted.

If a tag is a list, the extraction rule for the list will be applied iteratively to extract the list elements for that node in the structure. After that, other extraction rules will be applied to extract individual items like terms, compound terms, lexico-syntactic patterns, looking for subsumption, etc. As a final step, we use the structure to group together the individual items to assemble the thesaurus.

When we build the structure from an HTML document, the page is first split to a sequence of nodes. The extraction rules for each of the nodes in the structure are applied to the inner of tags until all tags have been consumed. After that, more rules are used to extract individual items.

6 Conclusion and future Work

We described a system for building a thesaurus from electronic documents and present results based on different combinations of techniques. The system serves as basis for a generalized framework for automated thesaurus construction tool.

The system is a semi-automatic thesaurus construction tool. It makes it easy to construct thesaurus by automating many of the tasks. The tool can be used for training new user on how to build the thesaurus. At the same time expert can use the tool to build his or her own thesaurus.

Different models for automatic thesaurus construction are presented in this work. Getting hierarchical relationships between terms entails an important advance in the process of thesaurus construction.

We addressed the problem how to effectively use the rich information, which is typically available in HTML to build thesaurus. We specified relevant hints about the hyperlinks and meta-data about web sites, which we hope, will aid in the task of thesaurus construction and aim at exploiting the richness of HTML. We examined these hints with different collection of web pages. We hope that our analysis will help future research into hypertext by providing some ideas about various types of information that may be present in a hypertext corpus and how one should take advantage of each.

The tool has the power for extracting the different sets of candidate terms and discovering the relationships between these terms. Although, the results achieved with our tool are quite encouraging. In most cases, it achieves an effective and accurate result. The tool needs manual constraints and the different models do not prevent from extraction of pairs which are not link by the target relation.

REFERENCES

1. R. Rada and B. Martin Augmenting Thesauri for Information Systems, ACM Transactions on Office Information Systems, 5 (4), 378--392 (1987).
2. ISO 5964, Documentation - Guidelines for the Development and Establishment of Multilingual Thesauri, 1985.
3. J. Aitchison, A. Gilchrist, and D. Bawden, Thesaurus Construction. 3ed. ASLIB, 1997.
4. J. Klavans, and E. Tzoukermann 'Combining corpus and machine-readable dictionary data for building bilingual lexicons' In Machine Translation, Vol.10, No.3, pp 1--34,1996.
5. M. A. Hearst, and H. Schutze, "Customizing a Lexicon to Better Suit a Computational Task", Proceedings of 31st Annual Meeting of ACL, Columbus, Ohio, USA, 55-69, 1993
6. P. Resnik, "Disambiguating Noun Groupings with respect to WordNet Senses", in S. Armstrong, K. Church, P. Isabelle, S. Manzi, E. Tzoukermann and D. Yarowsky (eds.), Natural Language Processing Using Very Large Corpora. Kluwer Academic Publishers, 1999, pp. 77-98. (Slightly revised version; the original paper appeared in the Proceedings of the 3rd Workshop on Very Large Corpora, MIT, 30 June 1995.
7 G. Salton, M. J. McGill, Introduction to modern information retrieval. McGraw Hill, New York. 1983.
8 Y. Abuzir, D. Vervenne, P. Kaczmarski and F. Vandamme "E-mail messages classification and user profiling by the use of semantic thesauri", Proceeding of CIDE2001 - 4e Colloque international sur le Document Électronique, Toulouse - France 24-26 octobre 2001.
9. E. M. Voorhees, "Using WordNet to Disambiguate Word Sense for Text Retrieval", Proc ACM SIGIR'93, Pittsburgh, 1993, 171-180.
10 Y. Abuzir "ThesConv: A tool for thesaurus construction from Prearranged list." to appear in proceedings of the 2002 International Conference on Information and Knowledge Engineeingr (IKE'02) part of a International Multiconferences in Computer Science, Monte Carlo Resort, Las Vegas, Nevada, USA, June 24 - 27, 2002.
11 H. Mili, R. Rada "Merging Thesauri: Principles and Evaluation". IEEE Transactions On Pattern Analysis and Machine Intelligence,10(2):204-220, 1988.
12. M.V. Mannino, S. B. Navathe, and W. "A Effelsberg , Rule-Based Approach for Merging Generalization Hierarchies". Information Systems, 13(3):257-272, 1988.
13. D. A. Evans, W. K. Ginther, M. Hart, R. G. Lefferts, and Monarch, I. A., Automatic indexing using selective nlp and first-order thesauri, In Proceedings of the RIAO, vol. 2, pp. 624-643, 1991.
14. G. Grefenstette, Use of syntatic context to produce term association lists for text retrieval. In SIGIR'92, pp. 89--97, 1992.
15. G. Ruge, Experiments on linguistically based term associations. In RIAO'91, pp. 528-545, 1991.
16. C. J. Crouch, An approach to the automatic construction of global thesauri, Information Processing & Management, 26(5): 629-40, 1990.
17. Y. Qiu, H.P. Frei, Concept Based Query Expansion. Proc. of the 16th Int. ACM SIGIR Conf. on R&D in Information Retrieval, Pittsburgh, SIGIR Forum, ACM Press, June 1993.
18. Güntzer, U., Jüttner, G., Seegmüller, G. and Sarre, F., Automatic Thesaurus Construction by Machine Learning from Retrieval Sessions. Information Processing & Management, Vol. 25, No. 3, pp. 265-273, 1989.
19. D. Roussinov and H. Chen, "A scalable self-organizing map algorithm for textual classification: A neural network approach to thesaurus generation." Communication Cognition and Artificial Intelligence, 1998.
20. Y. Abuzir, D. Vervenne, P. Kaczmarski and F. Vandamme, "TDOCS Thesauri for Concept-Based Document Retrieval", R&D Report BIKIT, BIKIT – LAE, University of Ghent 1999.
21. Hearst M. A., Automatic Acquisition of Hyponyms from Large Text Corpora. In Actes, 14th International Conference on Computational Linguistics (COLING'92), pages 539 545, Nantes, France, 1992.
22. Y. Abuzir, D. Vervenne, P. Kaczmarski and F. Vandamme, "Extracting Semantic Relationships between Terms using IKEM Tool", KIM/KIT NEWS , Vol. 15, nr.3, Nov. 2000.

Intl. Inform. & Libr. Rev. (2003), **35**, 383–397
doi:10.1016/S1057-2317(03)00033-X

A Knowledge Network Constructed by Integrating Classification, Thesaurus, and Metadata in Digital Library

JUN WANG*

ABSTRACT

For the digital libraries of China, the development of networked information resources, especially of metadata, is still a principal task. At the same time, many information resources have been accumulated and the exploitation of them is far from sufficient. The current chief utility — keyword search undermines seriously the potential value of information resources, especially the metadata, which contains valuable content description and indexing information. To solve this problem, we have devised a new paradigm — an integrated knowledge network. It is formed by merging the classification and thesaurus into a concept network and then distributing the metadata records into the concept nodes according to their subjects, just like the "shelving" of the metadata. We have built an experimental system, *VISION*, by incorporating a portion of the *Chinese Classification and Thesaurus* with the bibliographic data of the computing domain from Peking University Library. Such a knowledge network is not only a framework for metadata organizing, but also a structure for knowledge navigation, retrieval, and learning, and we believe it will be a catalyst for better knowledge management in digital libraries.

© 2003 Elsevier Ltd. All rights reserved.

INTRODUCTION

The library has been the center of the preservation, utilization, and distribution of information and knowledge of human beings for centuries. With the rising and the rapid development of the web, the role and function of the

* Information Management Department, Peking University, Peking, China. E-mail: junwang@pku.edu.cn

1057-2317/03/$ - see front matter © 2003 Elsevier Ltd. All rights reserved.

library are questioned. The web is the most widespread and easily accessible information resource, and its popularity continues to grow unabated. In turn, digital libraries have found a home on the web. Digital libraries distinguish themselves from other web information resources by having a much greater capacity for knowledge management. To digitize information and transfer it onto the web is far from knowledge management. How then do digital libraries achieve knowledge management? This is the problem confronting all libraries worldwide, including those in China.

Let's examine the China digital libraries more closely. On the one hand, the current development of networked information resources, especially metadata, is critical for Chinese digital libraries. For example, the China Digital Library Project (CDLP) and the China Academic Library and Information System (CALIS), the two biggest digital library projects in China, both spent a large proportion of their funds on the development of information resources.

On the other hand, many networked information resources have been accumulated. For example, in just 4 years CALIS has imported 89 databases containing 7800 kinds of academic journals, and its hundreds of members have produced 1150 thousand records for the union catalogue, 1370 thousand records of indexes from over 5500 Chinese academic journals, and 70,000 records of abstracts of theses and dissertations.[1] Compared with the production of resources, the utilization of them is deficient. Keyword-based searching is applied everywhere, both for indexed databases and full-text web pages. In the keyword matching, the valuable content description and indexing of the metadata, i.e., the subject descriptors and the classification notations, are merely treated as common keywords to be matched with the user query. Without the support of the vocabulary control tools (e.g. classification and thesaurus), content analysis, descriptive term searches, and metadata indexes are underused. New retrieval paradigms are needed in order to exploit the potentials of the metadata resources.

With these problems in mind, we review the structure of the traditional library. In order to provide high-quality document services for users, the documents are described and indexed first, which produces catalogues. The documents include books, journals, cassettes, and other items from the library. Then the catalogues are arranged in various categories, and the documents are organized into a given hierarchy (i.e., shelving). All these are done according to the classification and thesaurus. What is absent in current digital library architecture is the classification and thesaurus — the vocabulary control and the knowledge organization tools which serve three purposes in a traditional library: The description, organization, and

[1]Chen, L. (2001) The Report on the development of CALIS project at the end of first phrase (Chinese). http://www.calis.edu.cn/rm/calis211.pdf

retrieval of document collections. The first corresponds to metadata production. The second, including catalogue ordering and book shelving, tailors the general knowledge of classification and thesauri into the specific domains of the library collections, and incarnates them in a concrete way, viz., the ordered catalogues and organized shelves of books in library. On this basis, the readers can browse and search to see what they want.

For more efficient and effective exploration, the networked information should be pre-arranged together with vigorous improvement of search techniques, before the applicable information is made available on the web. Could classification and thesauri that contain the condensed intelligence of generations of librarians be used in digital libraries to organize the networked information, especially metadata, to facilitate the information resources usability and catalyze the digital library into knowledge management?

Yes. In light of the structure of traditional libraries, we design and implement a new paradigm which incorporates the classification, the thesaurus, and the metadata. The classification and the thesaurus are merged into a concept network, and the metadata are distributed into the nodes of the concept network according to their subjects. The abstract concept node substantiated with the related metadata records becomes a knowledge node. A coherent and consistent knowledge network is thus formed. It is not only a framework for resource organization, but also a structure for knowledge navigation, retrieval, and learning. We have built an experiment system based on the *Chinese Classification and Thesaurus*,[2] which is the most comprehensive and authoritative in China, and we have incorporated more than 5000 bibliographic records provided by Peking University Library. The results are indeed encouraging.

The next section presents a review on the classification and the thesaurus in China, and the section next presents the architecture of the knowledge network integrating classification, thesaurus, and metadata. In the penultimate section, the construction of our experimental system named *VISION* is explained step-by-step, and the final section is the conclusion and a look at the future.

A Review of the Classification and the Thesaurus in China

China, with its long literary tradition, has consistently used schemes of classification. Modern Chinese classification was greatly influenced by

[2]Liu, X.Sh., Zhang, Q.Y., *et al.* (Eds) (1994) *Chinese Classification and Thesaurus*, Beijing, Hua Yi Press. 1994.

Dewey, although the *Dewey Decimal Classification* wasn't popular in China. All the classifications now in use were created after the founding of the People's Republic of China. The most perfect one among them is the *Book Classification of Chinese Libraries* (BCCL) which was published first in 1975 and has undergone four revisions. Others worthy of mention are the *Book Classification of the People's University Library in China*, which is now in its sixth edition, and the *Book Classification of the China Science Academy Library*, which is now in its third version. There are also dozens of domain-specified classifications, but none of them is well-known.

The first Chinese thesaurus was produced in 1964. The most famous comprehensive thesaurus in China is the *Chinese Thesaurus* (CT), compiled by more than 1000 people over a 6 year period (1974–1980). It was the biggest thesaurus at that time, containing 91,158 preferred terms and 17,410 non-preferred terms. Influenced by the faceted thesaurus, a huge project was started in 1986 to combine the BCCL and the CT into one — the *Chinese Classification and Thesaurus* (CCT). More than 40 institutions were involved, and it was finished in 1994, with a total of 14 millions words in six volumes. Currently, it has been used in all the public libraries and more than 90 percent of non-public libraries and information institutions of China.[3] Now there are more than 100 various thesauri in China. Among them, the *Military Thesaurus of China* is well-known for being the only computerized thesaurus in China.

The CCT is not designed to be used in a network environment. To apply it in a digital library has several inherent obstacles; most of them are common to other classifications and thesauri.[4] The CCT was complied almost 10 years ago. Although it has undergone several amendments, it cannot keep up with the myriad variations in dynamic networked information. Two parts of the CCT, the classification and the thesaurus, are relatively independent of each other and cannot be updated synchronously. The CCT is a comprehensive system. Its general knowledge cannot be tailored flexibly into the specific domain of a given information collection; the CCT is a professional tool to be used by librarians. Its infrastructure, such as syntaxes and rules, are too complicated for common users to master. The CCT is mainly meant for the description and organization of information resources. It is not so good at retrieval; CCT's design is more oriented for the management of hard-copy documents, and some of its usage criteria lose their sense in the digital environments, such as one-place shelving (i.e. there is just one unique shelf place for a given book). For these

[3]Zhang, Q.Y., Liu, X.Sh. & Wang, D.B. (1996) Modern classification systems and thesauri in China. *62nd IFLA General Conference*, Booklet4, Beijing, August 25–31, 1996, pp. 31–37.

[4]Zhang, Q.Y. (1998) Discussions on the information retrieval language of 21 Century. *Forum of Libraries* **21**(5).

difficulties, there is seldom application of the CCT in the network environment.

THE KNOWLEDGE NETWORK

Classification, Thesaurus, and Bibliographic Data

It seems impossible to deploy traditional classifications and thesauri in digital libraries. But there exists one kind of networked information resource in which the classification and thesaurus are deployed thoroughly — the bibliographic data of the OPAC system. The bibliographic data is one of the most important resources the library (and only the library) owns. As collections of contemporary materials, they contain plenty of new professional terms in the title field, which supplement domain knowledge. More importantly, they are indexed strictly according to the classification and thesaurus. Based on the subject indexing, the bibliographic records can be combined with the classification and thesaurus to complement each other. It's a kind of metadata "shelving". The knowledge structure of the classification and thesaurus provides a skeleton for the organization of the bibliographic data; the concrete bibliographic data provide substance for the frame work. Thus, a corpus of knowledge is formed where new terms can be extracted automatically from the bibliographic data to update the classification and thesaurus; the classification and thesaurus is customized to the specific domains of the OPAC resources. Such a knowledge network provides the user with a natural structure for navigation, searching, and learning. For convenience, we designate such knowledge network as "KNICTM" (Knowledge Network Integrating the Classification, Thesaurus, and Metadata).

The Construction of the Knowledge Network

The KNICTM is built in three steps:

- *Construction of the original concept node based on the classification and thesaurus.* First, the thesaurus is turned into an original concept network that consists of nodes and edges. A node is composed of the synonym set of a subject descriptor, including the descriptor and all the terms connected to it by the equivalent relationship in the thesaurus. If there is a hierarchical relationship between two terms, an "is-a" edge is set up between the corresponding concept nodes. The associative relationship is discussed in the next step. Then the classification scheme is embedded into this original concept network, and acts as a disciplined hierarchical backbone for the concept network (Fig. 1). The CCT is a reciprocal index between the BCCL and the CT rather than a faceted thesaurus, and there is no direct mapping between the categories of

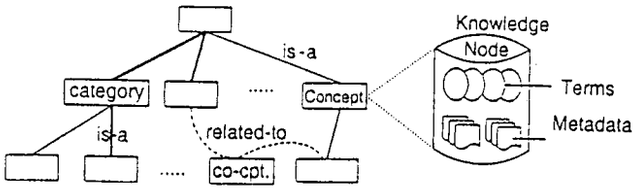


FIGURE 1. The knowledge network with a close-up of a knowledge node.

the BCCTand the concepts of the CCT. Therefore, the category nodes are created and the relationships between the category nodes and the concept nodes are set up.

- *Distribution of the bibliographic data to the concept network.* Secondly, the bibliographic data are distributed into the nodes of the original concept network according to their subjects. This is the key task of the KNICTM construction. Bound with the corresponding bibliographic data records, an abstract concept node becomes a knowledge node (Fig. 1) and the concept network turns into a knowledge network. The distribution of the bibliographic data records (BDR) in the original concept network is explained as follows: If a BDR contains only one subject descriptor, we take the BDR as one of the instances of the corresponding concept node and add the BDR into the node. If a BDR contains several descriptors, then we add the BDR into all the related concept nodes as instances of them. If a BDR contains a composite subject which is denoted by a coordination of descriptors, then we create a new concept node, and connect it to all the corresponding nodes of the coordinate descriptors with "related-to" edges. The newly-created concept node is called a "co-concept" node, which has the BDRs merely and no term for the moment. For example, a BDR with the title "*Internet Firewall Technologies*" is indexed with the string "Network-Security", for there is no "firewall" in the thesaurus. To add this BDR into the concept network, a co-concept node is created, and it is connected to the concept node of "Network" and the concept node of "Security" by 'related-to' edges. Because the associative relationship easily goes out of control in the thesaurus, we don't consider it in the first step. Only when the correlation of two concepts is supported by a BDR do we establish the 'related-to' relationship between them through a co-concept. Thus, the BDRs verify the associative relationship. In the next step, when there is a new extracted term with the meaning of the co-concept, the new term is added to the co-concept. In the above-given example, this is a 'firewall'. The KNICTM needs manual examination periodically to confirm the co-concepts created.

When a preferred term is determined for the co-concept, the co-concept node turns into a common concept node.

- *Enhancement of the KNICTM.* The last and the most difficult task is to mine new terms from the metadata collection to enhance the KNICTM. The title of a scientific document usually summarizes its content and reveals its central topics. And, there exists a direct mapping between the keywords of the title and the subject indexing. Based on their semantic mapping, new terms can be extracted from the title and added into the concept network. There are three difficulties to accomplishing this. How to extract valuable terms out of the common terms in title? How to determine the position where the extracted terms should be inserted into the KNICTM? Prior to all these processes, the titles of natural language must be segmented into word or phrase, the classical obstacle of Chinese language processing.

The construction of the KNICTM is similar to a tree. The classification hierarchy is the trunk and branches of the tree; the conceptual relationships of the thesaurus are the veins, and the metadata are the leaves. The basic construction unit of the KNICTM is the knowledge node. It is the concept node attached with references. Two kinds of edges join the knowledge nodes together: the hierarchical edges ("is-a"), which are the meridians of longitude of the KNICTM, and the associative edges ("related-to"), which are the parallels of latitude.

The Functions of KNICTM

- *A Framework for the organization of network resources.* The KNICTM provides a framework for the organization of the networked information resources, especially metadata. It is a network of knowledge with substantial data appended rather than a mere abstract concept network. As the instances of the concept, the metadata records inherit all the relationships among the concepts. The records of metadata which were isolated from each other become semantically connected now and are woven into an interconnected knowledge network.
- *An adaptive concept network based on the applied resources.* The classification and thesaurus represent general knowledge, and cannot fit a specific information collection perfectly. The KNICTM is an adaptive concept network capable of self-customizing, based on the scale and domain of the given collection. The nodes and edges supported by the metadata instances prove the usability of the correspondent concepts

and relationships, and they are revealed to the user. The nodes and edges without the support of the metadata instances indicate that the correspondent concepts and relationships are unusable and may need updating. Furthermore, statistic and semantic techniques can be applied in mining new terms, concepts, and relationships from metadata collection to enrich concept networks automatically.

- *A structure for knowledge navigation and retrieval.* The keyword-based search undermines the value of the metadata seriously. The KNICTM provides a conceptual retrieval network and visual navigational ontology. First, the KNICTM can guide the user to clarify his information demands and express his query clearly. Second, because all the metadata have been arranged into the KNICTM, there is no need to dig into the metadata collection by keyword searching; it is only necessary to locate the knowledge node according to the user query, and follow the surrounding edges to reach other nodes to obtain the desired result. Third, now that all the metadata have been arranged into the structure of the KNICTM according to their subjects, the retrieval result is displayed in that structure, already ranked and classified.
- *A well-organized knowledge network to support knowledge learning.* The KNICTM consists of the knowledge node that is identified by a concept composed of a set of synonymous terms and supported by some metadata records. The knowledge nodes are organized into a hierarchy and clustered into topic areas through edges among them. A friendly interface like Cat-a-Core[5] can manifest the structure of the knowledge network. A user assisted by such an interface can learn the discipline structure of a domain, master the professional terms, understand the relationships among the subjects, and pick up the documents to study.
- *A digital library of knowledge management.* The most essential element of a library is information. In KNICTM, information resources are integrated into a knowledge structure described by both classification and thesaurus. The KNICTM can be easily extended to support other activities of a digital library, such as resource collecting and indexing. Thus, all the activities of a library, including indexing, organization, navigation, retrieval, and learning, center on this knowledge network. To develop continuously, the KNICTM must move the digital library from information management to knowledge management.

[5]Hearst, M.A. (1997) Cat-a-cone: An interactive interface for specifying searches and viewing retrieval results using a large category hierarchy. *Proceedings of the Twentieth Annual International ACM SIGIR Conference*, Philadelphia, PA, July 1997.

THE BUILDING OF AN EXPERIMENT: THE VISION SYSTEM

We built an experimental system named *VISION*. In *VISION*, we combined the classification and indexing terms from the computing domain in the CCT with all the bibliographic records for Chinese computer science materials held by the Peking University Library, published between 1990 and 1999, totaling 5324 records. The *VISION* system has a client/server architecure; on the server side is the knowledge network supported by Oracle9i; on the client side is the user interface implemented in Java. We chose Oracle9i, for we need its powerful object-oriented feature such as nested table and variable array to support our complex objects. Java was selected in order to make it easy to transplant the system to the web.

We will first introduce the ontology of our system. Then we outline the architecture of the system on the server side and the function of the conceptual retrieval tool on the client side. Finally, we will introduce the work we have done in extracting new terms from the bibliographic data to enhance the knowledge network.

The Ontology Design

There are many objects in our system and their relationships are complex, so we used ontology tools such as Ontolingua[6] and Protégé[7] to design our system's ontology. Then, we converted this ontology into the database schema. Our ontology consists of seven classes: term, concept, co-concept, category, document, author, and publisher. Fig. 2 depicts their relationships, with the numbers indicating the cardinality of the links. Pieces of ontology in the syntax of Ontolingua and in the RDF schema rendered by Protégé are provided in Appendices A and B for further reference. We then convert the ontology into the relational schema of the database system and create the corresponding tables in Oracle.

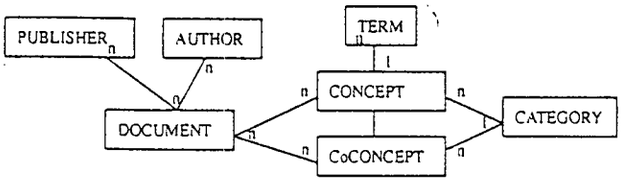


FIGURE 2. The objects and the relationships in the *VISION* system.

[6]Gruber, T.R. (1993) A translation approach to portable ontology specifications, *Knowledge Acquisition* 5, pp. 199–220.

[7]Noy, N., Fergerson, R. & Musen, R. (2000) The knowledge model of protege-2000: Combining interoperability and flexibility, http://www.smi.stanford.edu/pubs/SMI_Reports/SMI-2000-0830.pdf

VISION on the Server Side: the Knowledge Network

The original data set of the *VISION* system includes the e-text of the CCT and the bibliographic data of the computing domain. Both of these were provided by the Peking University Library. The characteristics of the original data have considerable influences on the system design and implementation. There are three steps to building the *VISION* system on the server side:

- *The e-text file of the CCT is processed to set up the fundamental structure of the VISION* system. A particular tool was developed to serve this purpose. The e-text of the CCT is read in, and all the entries (i.e. categories and terms) on computer science are processed. According to the structure, layout, and notation rules of these entries, the related records are created and appended into four tables, respectively: TERM, CONCEPT, CoCONCEPT, and CATEGORY. In this process, we get 2194 terms, including 1684 preferred terms and 278 non-preferred terms; and 1684 concepts are created. Some non-computing-domain terms are also captured, for they are the related terms of the computing domain terms. Every subject is treated as a concept, so the quantity of the preferred terms is equal to the quantity of the concepts.
- The bibliographic data are loaded in the database and distributed into the original concept network constructed in the preceding process. The bibliographic data are in CNMARC format. A particular tool is developed to decode the perplex CNMARC format and read each bibliographic record out, and the required fields are extracted (title, subject, author, etc.) to form a new record, which is appended into the table DOCUMENT. A total of 5053 document records have been created. Others are discarded for various reasons, for example, unrecognized titles or dual ISBN numbers. Such data processing requires a lot of time and energy. After the data loading, the records of the DOCUMENT table are connected with the records of the CONCEPT table, based on the semantic correspondence between the subject indexing of the document records and the concept records. When necessary, new records are created in the CoCONCEPT table.
- *New terms are extracted from the DOCUMENT table and added to enhance the knowledge network of the VISION* system. Some of the problems have already been mentioned. This process is the focus of the ongoing second phase of the Vision project, so here we outline roughly what we have done to date.

(a) *The extraction:* At present a statistic algorithm is applied to extract terms in the titles. First, the title is segmented into basic words and phrases by general segmentation tools, and then the co-occurrence

frequency of the neighbors is counted. If the frequency is bigger than a given threshold, the combination is selected as a new-term candidate τ; then all the indexing terms are collected from the document records whose title field contain τ. Thus a set of indexing terms I is formed. The distribution of I is computed. If the distribution of I is convergent, then τ is a new term.

(b) *The insertion:* The convergent point found above helps to determine the position where the new term should be inserted. Generally, the extracted new term is the narrower term of the subject represented by the convergent point. Now we are considering applying Lattice Theory[8] or Formal Concept Analysis[9] to this problem.

The VISION System on the Client Side: Knowledge Navigation and Retrieval

We implemented a knowledge retrieval system in Java to navigate and retrieve the *VISION* knowledge network. Fig. 3 is the snap-shot of the user interface. There are four physical areas in the interface: the query dialog, the concept network window, the information window, and the document window. Within the concept window, there are three basic ways to view the concept network: hierarchical tree, alphabetical list, and concept family. The hierarchical tree is similar to a faceted thesaurus. All the categories and concepts (identified by the preferred terms) are organized into an expandable conceptual tree. The alphabetical list is an index of all the terms in alphabetical (Chinese Pin Yin) order. The concepts can also be organized into concept families, that is, the term families of the thesaurus, and listed in alphabetical order by the top concepts. There is a fourth option, which is a hybrid of the hierarchical tree and the alphabetical list. When the user clicks on a concept, its detail information is displayed in the information window, including its term set, super-concept, sub-concepts, corresponding category, and the co-concepts around it; the documents connected with it are displayed in the document window. All the windows trigger each other and act in a chain, and all the objects in the windows are clickable. When a query is submitted, what is returned are not only the documents related, but also the concepts and co-concepts pertinent. The concept network window is triggered to highlight the position of the retrieved concepts in the concept hierarchy, and the relationships with other concepts are shown; the document window is triggered and the relevant documents are revealed. All these come to the user together as an integral knowledge unit.

[8]Pedersen, G.S. (1993) A browser for bibliographic information retrieval, based on an application of lattice theory. *ACM-SIG'93-6/93/Pittsburgh*, PA, USA.

[9]Ganter, B. (1999). *Formal concept analysis*. Berlin, Heidelberg: Springer.

critical for the concept network to be self-sufficient. Now we consider applying an adjusted Formal Concept Analysis to this problem.

- A language for the concept query and manipulation in the concept network will simplify the operation of the concept network and add the automatic query expansion and contraction mechanism.
- A visualization interface such as Cat-a-Core[10] or Inxight Star Tree[11] can provide more friendly interaction with the user. A visualization interface that is structurally isomorphic to the concept network will support knowledge learning more powerfully.

When all the aspects of the system have been tested, the system will be transplanted to the web and incorporated with the current OPAC system of Peking University.

To integrate the classification, thesaurus, and metadata into a coherent knowledge network has promising applications in digital libraries. It could easily be expanded to support automatic classification and indexing in scientific domains. Enhanced by the bibliographic data, the knowledge network could absorb other metadata, such as index databases of journals, magazines, or newspapers. The web community also recognized the importance of the standardization and organization of the web information. XML, RDF, Dublin Core, and other specifications are preparing the web for the manageable web — the Semantic Web as envisioned by Tim Berners-Lee.[10] But how to construct it? Our paradigm provides one approach.

[10]See footnote 5.

[11]Inxight Software Inc. (2001), Inxight Software, http://www.inxight.com

ACKNOWLEDGEMENTS

This research is an excerpt from my doctoral dissertation written at Peking University. I wish particularly to thank Zhu Xingguo, Deng Peng, and Zu Yong, all of whom assisted me in developing the Vision system. In addition, I would like to thank Professors Yang Dongqin and Tang Shiwei, as well as Dai Longji, President of the Peking University Library, and his staff.

APPENDIX A. VISION'S ONTOLOGY IN ONTOLINGUA

```
(define-class TERM (?x)
  :def (and (ctms-thing) ?x))
  :axiom-def (subclass-partition
                TERM (setof preferred-term non-preferred-term)))
(define-class CONCEPT (?x)
  :def (and (ctms-thing) ?x)
            (has-one ?x cpt.identifier))
(define-function CPT.IDENTIFIER (?cpt) :-> ?identifier
  :def (and (concept ?cpt)
            (preferred-term ?identifier)))
(define-relation CPT.TERMS (?cpt ?terms)
  :def (and (concept ?cpt)
            (term ?terms))))
(define-relation CPT.CATEGORY (?cpt ?ctg)
  :def (and (concept ?cpt)              (category ?ctg))))
(define-relation CPT.BROADER-CONCEPT (?cpt ?b-cpt)
  :def (and (concept ?cpt)
            (concept ?b-cpt))))
(define-relation CPT.NARROWER-CONCEPT (?cpt ?n-cpt)
  :def (and (concept ?cpt)
            (concept ?n-cpt))))
(define-relation CPT.DOC-CONTAINED(?cpt ?docs)
  :def (and (concept ?cpt)
            (document ?docs))))
```

APPENDIX B. VISION'S ONTOLOGY IN RDFS (PROTÉGÉ)

```
<?xml version='1.0' encoding='ISO-8859-1'?>
<!DOCTYPE rdf:RDF [
    <!ENTITY rdf 'http://www.w3.org/1999/02/22-rdf-syntax-ns#'>
    <!ENTITY thesaurus 'http://protege.stanford.edu/thesaurus#'>
    <!ENTITY rdfs 'http://www.w3.org/TR/1999/PR-rdf-schema-19990303#'>
]>
<rdf:RDF xmlns:rdf="&rdf;"
    xmlns:thesaurus="&thesaurus;"
    xmlns:rdfs="&rdfs;">
<rdfs:Class rdf:about="&thesaurus;*ConceptFamily"
    rdfs:label="*ConceptFamily">
   <rdfs:subClassOf rdf:resource="&rdfs;Resource"/>
</rdfs:Class>
<rdfs:Class rdf:about="&thesaurus;?NewTerm"
    rdfs:label="?NewTerm">
   <rdfs:subClassOf rdf:resource="&thesaurus;Term"/>
</rdfs:Class>
<rdfs:Class rdf:about="&thesaurus;Book"
    rdfs:label="Book">
   <rdfs:subClassOf rdf:resource="&thesaurus;Document"/>
</rdfs:Class>
<rdfs:Class rdf:about="&thesaurus;Category"
    rdfs:label="Category">
   <rdfs:subClassOf rdf:resource="&rdfs;Resource"/>
</rdfs:Class>
<rdfs:Class rdf:about="&thesaurus;Category_Code"
    rdfs:label="Category_Code">
   <rdfs:subClassOf rdf:resource="&rdfs;Resource"/>
</rdfs:Class>
<rdfs:Class rdf:about="&thesaurus;Concept"
    rdfs:label="Concept">
   <rdfs:subClassOf rdf:resource="&rdfs;Resource"/>
</rdfs:Class>
<rdfs:Class rdf:about="&thesaurus;Conference"
    rdfs:label="Conference">
   <rdfs:subClassOf rdf:resource="&thesaurus;Event"/>
</rdfs:Class>
```

Pergamon

Information Processing & Management, Vol. 33, No. 4, pp. 481–493, 1997
© 1997 Elsevier Science Ltd
All rights reserved. Printed in Great Britain
0306-4573/97 $17+0.00

PII: S0306-4573(97)00009-5

THESAURUS CONSTRUCTION: PROBLEMS AND THEIR ROOTS

URI MILLER

Senior Indexer, Library, Wingate Institute for Physical Education and Sport, Wingate P.O., Netanya, 42902, Israel, Thesauri Construction Coordinator, MOFET Institute—Research and Development of Programmes for Teacher Education, Ministry of Education, Culture and Sport of Israel.

(Received 15 November 1996; accepted 20 February 1997)

Abstract—Some general problems of thesaurus construction theory and practice are analyzed and discussed. The analysis is based on research and experience already accumulated in librarianship and information science as well as some other sciences and general scientific methods (linguistics, systems analysis, etc.). Some methodological and methodical ways of solving problems existing in thesaurus construction are recommended.
© 1997 Elsevier Science Ltd

1. INTRODUCTION

> What has happened will happen again, and what has been done will be done again, and there is nothing new under the sun. Is there anything of which one can say, 'Look, this is new?' No, it has already existed, long ago before our time. The men of old are not remembered, and those who follow will not be remembered by those who follow them. Ecclesiastes 1:9-12

The research process in the field of library and information science has, among others, one remarkable feature: the upsurges of interest in one theme or another are characterized by a certain periodicity which seems to depend primarily on the penetration and introduction of hi-tech progress in this sphere. Indeed, we can see that three fundamental innovations—first, the general computerization of our libraries; second, the appearance of searching tools on optical discs in their reading halls and reference departments; and, third, the wildly enthusiastic fall of all world information community into the nets of Internet—became the reasons not only for change and development of different library mechanisms, but also for general analysis of their construction methods and of their influence on library and database processes, as well as for elaboration of different ways of information storage and retrieval which seem optima to their creators. As was already noted, 'all the future exists in the past'—and therefore just now, amidst the boom of CD-ROM databases, multimedias and the Internet (which, on the one hand, can really open broad-scaled opportunities not only in the nearest future but now, on the other hand, can inspire, as has already happened, some unjustified illusions) it seems especially timely to discuss certain principal methodological aspects of thesaurus construction, to summarize already accumulated experience and to try to determine some perspectives of this process[1].

This article may seem to be a kind of literature review, but every analysis of some experience is based on the concrete results, expressed, first of all, in printed form, and I hope for forgiveness for sometimes excessive references. After all, current knowledge is primarily based on the achievements of our predecessors[2].

[1]Such a need was emphasized by Milstead (1994).

[2]See also Nedobity (1983); Beghtol (1986); Vickery (1986); Teskey (1989); Hjørland (1992); some of my propositions were previously formulated for the traditional information retrieval language of subject headings in Miller (1980).

Supplied by the British Library - "The world's knowledge" www.bl.uk

In this connection the following have been noted:

1. the steady and indomitable growth of information inevitably leads to impossibility of its proper processing and search by manual methods;
2. the already proven ability of librarians and information workers to filter the appropriate materials (see, for example, Kuller *et al.*, 1993) allows to assert that the veritable degree of every electronic database's completeness may be verified by the comparison of documents received by a manual search but missed by a electronic search with their presentation in citation indexes which can give us a real picture of indexing quality; and
3. the results of tests confirming the greater completeness of manual search, in my opinion, first of all demonstrate not the superiority of manual search but the defects of the concrete database, mainly, its indexing and searching tools, first and foremost, thesauri. (It is precisely this fact that explains the test's results of Reich and Biever, 1991.)

We can add one more consideration of a psychological nature. Information storage and retrieval are the main parts of the concept 'information service', and the foregoing opposition seems to be the refraction of the general attitude to the service problems. The fear of an irrepressibly growing stream of information together with lack of self-confidence and ability to manage this stream gave rise to a yearning to set up a certain informational 'self-service store' functioning by the principle 'Take what you want!' But even such a service urgently needs specific organization of its functional process, and the assumption that such process can be self-managed is an obvious delusion.

Of course, these debates did not end—and in essence could not end—with a victory of one of the opposing groups, but they permitted drawing some conclusions about weak spots of thesauri. Briefly this weakness was formulated by Milstead (1992) in the following way: 'the danger is that if the thesaurus is permitted to become monolithic and resistant to change, it can actually hinder both indexing and retrieval'.

3. THESAURUS CONSTRUCTION: SOME ELEMENTS OF THEORY

3.1. Natural language or conceptual network?

The systematization and regulation of scientific knowledge, converted into scientific information, is determined not only by an object of cognition but also by a subject of cognition, a system of knowledge itself, traditions and social peculiarities which exist in every concrete scientific activity. This process is obviously the reflecting one, but it is 'a reflection of the second degree', or 'reflection of reflection' because it reflects not only an object itself but a subject's knowledge and cultural system.

A process of interpersonal communications and knowledge exchange is realized by means of signs, including words, which have a multiplicity of meanings[9]. A meaning is certain information about an object or a concept. Every concrete meaning of a sign is defined by a concrete context of this sign's usage. Signs as a means of information transfer have a conventional nature, and this feature gains strength from perceptual conceptions to elements of natural languages and reaches its culmination in the elements of artificial languages. Their conventional nature is determined by social factors as well as by system connections of signs, and in this sense a sign depends on conceptual reality. Thus a meaning is not a direct reflection of reality but serves as an auxiliary aspect of reflection for its development as well as for an exchange of its results in interpersonal communications.

By contrast to contextually dependent words, terms are signs, or lexical symbols, with one or several strictly definite and distinct meanings which are crystallized from the whole lexical

[9]Many interesting observations in connection to this problem are collected in Seiler and Wannemacher (1983); see also Yngve (1981); Blair (1990); Dahlberg (1992); Ingwersen (1992); Hjørland (1994); Gillaspie (1995); Sharada (1995).

enrichment of the indexing vocabulary (see Jamieson *et al.*, 1986; Gomez *et al.*, 1990), i.e. in particular the maximal saturation of the indexing vocabulary by synonyms for descriptors, can be the resolution of the second contradiction. This decision is not only in conformity with the concept of 'relevance' studied by Schamber *et al.* (1990), because it helps users more quickly and properly to find *their own* relevant documents, but is very closed to the interesting method of 'semantic hyperindexing' discussed by Arents and Bogaerts (1993), Pollard (1993) and some other researchers.

Therefore *the structure of an information retrieval language simulates not a system of natural language but a system of concepts, 'a conceptual network'.*

3.2. Systems approach: the unique tool

Simulation in its essence implies *the systems approach* to an object, and thesaurus construction is not an exception to this rule. The expediency of application of systems analysis methods to the sphere of information science has been known for a long time[12], therefore I want only to mention the focal points which are significant for our topic.

To begin with, it is necessary to note that 'system thinking is first and foremost a point of view and a methodology arising out of this viewpoint' (Mattessich, 1982, p. 384). This kind of thinking, 'the science of simplification' (as W. R. Ashby called systems analysis), gives the opportunity to reveal cardinal and typical features of every system:

1. its integrated character, determined, apart from everything else, by system elements and its environment;
2. existence, among others, of basic steady (system-forming) relations between its elements;
3. existence of a structure which is based on these relations, which provides system streamlining and organization, and which may be characterized both horizontally (concepts of the same level) and vertically (concepts of different levels and their hierarchy);
4. existence of relations between different levels (system management);
5. feedback between the system and its environment which ensures constant system modification.

Therefore the optimal thesaurus must reflect the system essence of problems as well as be a system itself.

The systems approach (as a microscope) gives the possibility to create a model with different degrees of approximation (from the most detailed for the central, 'nuclear', concepts to the most generalized for the peripheral spheres) but at the same time to keep the system integrity.

Every language possesses its specific *term system* with a vocabulary as its real embodiment, where its specific terms coexist with general terms and terms from other semantic fields. In this connection the following can be cited: 'Although thesauri differ in detail they all share a basic principle: they record a set of terms (words of phrases) covering some knowledge domain, and three types of relationship (equivalence, hierarchical and associative) between them...' (Jones, 1993)[13].

Finally, only the systems approach can be the methodological basis of new generations of information retrieval systems (built, for example, in 'virtual reality' (Newby, 1993) because of the immanent systems essence of conceptual space modelled by them[14]. It is important to take into consideration that the very interesting and very promising concept of 'navigation' is in point of fact the rules of *orienteering* in a hi-tech information system built on the basis of the systems approach[15], and what is more, such kinds of thesaurus design as 'graphical', as was also noted

[12]See, for example, Vilenskaya (1969, 1975); Borko (1980); Miller (1980); Mansfield (1982); Mattessich (1982); Rao and Zunde (1982); Cochrane (1984); Lawrence (1985); Iyer (1992); Rubin (1992), etc.

[13]See also Dimitrova (1992); very interesting studies of these relations in the aspect of the problem of relevance see Green (1995a,b); Green and Bean (1995).

[14]Some examples of consequences of the ignorance of system essence of the concrete conceptual network in the functioning thesaurus are demonstrated in Miller (1996).

[15]See, for example, Newby (1991), (1992); Buckland *et al.* (1995), etc.

Supplied by the British Library - "The world's knowledge" www.bl.uk

But, as Bruner (1974) ascertains, 'A coding system may be defined as a set of contingently related, nonspecific categories. It is a person's manner of grouping and relating information about his world, and it is constantly subject to change and reorganization.' Therefore one and the same concept can be simultaneously ascribed to different categories of one and the same categorial system depending on the aspect under examination (see: Körner, 1970)[19].

This property is the specific one which distinguishes categorial analysis from cluster analysis already adopted for library research[20]. Generally speaking, the difference between the two *logico-semantical* methods consists in their nature because: (a) categorial analysis is based on categories constructed *beforehand* but clusters are formed *during* this analytical process, and (b) every term can be simultaneously included in *some* categories but only in *one* cluster. Therefore we can consider categorial analysis as a thesaural method and cluster analysis a classification method (see, for example, Larson, 1991b). Theoretically, and even practically, it is possible to build a separate hierarchy of 'genus–species' relations proceeding from every aspect of a concept, and their aggregate must coincide with a 'genus–species' hierarchical structure of descriptor as a concept's vocabulary embodiment. (Exactly this idea has been realized in the MEDLINE database on the optical disc, where an user can search a material from thesaurus by means of some 'hierarchical trees'). The vital importance of polyhierarchy was perfectly described already many years ago by Soergel (1974). It must be noted that in practice we can quite often encounter absolute misunderstanding of this thesaurus specificity expressed in the need for the presence for every descriptor of only one genus-related term, though some authors have with some reluctance admitted polyhierarchy but only as a rare exception to the rule. For illustration Batty (1989) can be cited, who wrote: 'Clusters of terms should be mutually exclusive, as far as possible, that is, no term in one cluster should appear in any other clusters without good reason. Sometimes they may: the term ASPIRIN may appear in both ORGANIC CHEMICALS, as a SALICYLIC ACID, and in DEPRESSANT DRUGS. ASPIRIN, then, has two BT (Broader Term) relationships: to the higher term in ORGANIC CHEMICALS and to the higher term in DRUGS. This is known as a polyhierarchy, something that lexicographers try to avoid if possible, but accept if inevitable.'

Even Aitchison (1982) names polyhierarchical relations of descriptor only as *ones of* relations, but not as ones which **determine** thesauri's specificity. But elsewhere the same author (Aitchison, 1986) asserts: 'The alphabetically arranged thesaurus... has as its base a classificatory structure which is not immediately apparent... a powerful tool for organizing the subject field is facet analysis, which engenders a more precise understanding of the terms and how they are interrelated. In facet analysis, a systematic schedule or faceted classification is constructed from which is later derived the conventional alphabetical thesaurus'. In this connection extremely significant is the way of solution by Aitchison of the problem of 'the repetition of identical concepts in several places in the schedules', i.e. the multiaspectual essence of a concept and, hence, a term. The author proposes the following solution: 'While the same thesaurus term could be linked to more than one class number, thesaurus compilation is simplified if a preferred place is selected for the concept in the schedules and a cross-reference made, in the form of hierarchical or associative relationships between the preferred and non-preferred location' (*Ibid.*).

It appears that such an expression clearly shows an absolutely classification-based way of thinking because 'a preferred place' can be selected only in classification schemes but not in alphabetically arranged thesauri. Besides, here we deal with very strange meaning of the thesauri-constructing term 'cross-reference' as a form of 'hierarchical or associative relationships between preferred and non-preferred location'. It is obvious that Aitchison speaks about descriptors but means in this case only location in classification schedules because the relationships between preferred and non-preferred terms are not hierarchical or associative but only the relations of equivalence, and there is only one conventional way of expressing these relations between preferred descriptor and non-preferred terms—references which have not

[19]For very interesting thoughts about categorization and its role in information storage and retrieval see Case (1991); Warner and Wenzel (1991); Hjørland (1994); Jacob (1994). See also Vickery and Vickery (1992).

[20]See, for example, Stanfel (1983); Willett (1988); Gordon (1991), etc.

Supplied by the British Library - "The world's knowledge" www.bl.uk

To my thinking, a thesaurus may be defined as *a lexico-semantical model of a conceptual reality or its constituent, which is expressed in the form of a system of terms and their relations, offers access via multiple aspects and is used as a processing and searching tool of an information retrieval unit*. This point of view on thesaurus gives us an adequate interpretation of realities and brings the subject and the object of information processing and retrieval into proper correlation.

It is clear from this definition that there is no place for any attempt at the functional division of the thesaurus into an 'indexer thesaurus' and an 'end-user thesaurus' even within the framework of a single system, as some researchers propose[23]. The reasons are obvious. Such an idea:

1. contradicts the systems approach because users are just that 'system environment',
2. wrecks the system unity and breaks one of the fundamental system rules—permanent intercommunications with the system environment,
3. reduces user possibilities for access to information because of artificial difficulties in understanding the indexer's logic of activities as well as indexer control of feasible ways of user access to information.

Hence it appears that a process of thesaurus construction is a process of simulation in a lexical form:

1. of the whole universum of realities and concepts or its part, and
2. of hierarchical and associative connections and relations between these realities and concepts.

5. THESAURUS ROLE IN INFORMATION STORAGE AND RETRIEVAL

Here we must take into account that on the basis of the existing information materials many practical information workers make a sharp and even essential distinction and contrast between two roles of the thesaurus—theoretical (as a model) and practical (as a real tool for information indexing and retrieval). Such interpretation seems to be a reflection of:

1. misunderstanding of precisely that thesaurus essence which I tried to show, and
2. psychological problems connected with the whole complex of information service and expressed sometimes in blind idolization of a user.

As to users and their really and undoubtedly central place in the information service, I want to cite the words of one of the most prominent specialists—Austin (1986). He considers as obvious the "fact that we cannot, with the best will in the world, construct an indexing language that matches every expectation of every user. In cases where it appears that a choice must be made between an indexing language that some users would favour because it is 'stylish', and another that would be rated as good because it works, I suggest we should choose the latter and stand by that decision".

In my opinion, a methodologically correctly constructed thesaurus gives certain essentially significant advantages to the whole information retrieval system which it supports.

Firstly, a thesaurus, constructed on the system principle, qualitatively increases 'hospitality' to new concepts because the process of semantical simulation is the 'mirror' of the cognitive processes where every new concept receives in any case its own proper place in the conceptual network.

Secondly (and this statement is inseparable from the first), such thesaurus construction philosophy sharply reduces even the possibility of divergences of conceptual interpretation of one and the same descriptor by:

1. two different indexers,
2. one and the same indexer at different times,

[23]See, for example, Bates (1986); Reich and Biever (1991); Morris (1994).

DeNeef, P. (1988). The comprehensiveness of computer-assisted searches of the medical literature. *J. Family Practice, 27*, 404–408.
Dimitrova, G. (1992). Thesaurus semiotics and thesaurus lexicography. *International Classification, 19*(2), 86
Dubois, C. P. R. (1987). Free text vs controlled vocabulary: a reassessment. *Online Review, 11*, 243–253.
Fidel, R. (1987). Controlled vocabulary and free-text searching: searcher's selection of search keys. *Proc. ASIS Annual Meeting* (Vol. 24, pp. 71–73).
Fidel, R. (1988). Factors affecting the selection of search keys. *Proc. ASIS Annual Meeting* (Vol. 25, pp. 76–79).
Fidel, R. (1991). Searchers' selection of search keys: 1. The selection routine. 2. Controlled vocabulary of free-text searching. 3 Searching styles. *J. Ameri. Soc. Info. Sci., 42*, 490–527.
Fidel, R. (1992). Who needs controlled vocabulary?. *Special Libraries, 83*(1), 1–9.
Foskett, D. J. (1980). Thesaurus. In A. Kent (Ed.), *Encyclopedia of library and information science* (Vol. 30, pp. 416–463). N.Y.: Marcel Dekker, Inc.
Gillaspie, D. L. (1995). The role of linguistic phenomena in retrieval performance. *Proc. ASIS Annual Meeting* (Vol. 32, pp. 90–96).
Gomez, L. M., Lochbaum, C. L., & Landauer, T. K. (1990). All the right words: finding what you want as a function of richness of indexing vocabulary. *J. Amer. Soc. Info. Sci., 41*, 547–559.
Gordon, M. D. (1991). User-based document clustering by redescribing subject descriptions with a genetic algorithm. *J. Amer. Soc. Info. Sci., 42*, 311–322.
Green, R. (1995a). Syntagmatic relationships in index languages: a reassessment. *Library Quarterly, 65*, 365–385.
Green, R. (1995b). Topical relevance relationships. I. Why topic matching fails. *J. Amer. Soc. Info. Sci., 46*, 646–653.
Green, R., & Bean, C. A. (1995). Topical relevance relationships. II. An exploratory study and preliminary typology. *J. Amer. Soc. Info. Sci., 46*, 654–662.
Gregor, D., & Mandel, C. (1991). Cataloging must change! *Library Journal, 116*(6), 42–47.
Hjelmslev, L. (1970). *Language: An Introduction*. Madison, WI.-London: Univ. of Wisconsin Press.
Hjørland, B. (1992). The concept of 'subject' in information science. *J. Documentation, 48*, 172–200.
Hjørland, B. (1994). Nine principles of knowledge organization. In H. Albrechtsen, S. Oernager (Ed.), *Knowledge Organization and Quality Management: Proceedings of the Third Int. ISKO Conf.*, 20–24 June 1994, Copenhagen, Denmark (Advances in knowledge organization, Vol. 4, pp. 91–100). Frankfurt-Maine: Indeks Verlag.
Hodges, P. R. (1983). Keyword in title indexes. *Special Libraries, 74*(1), 56–60.
Hornby, A. S. (1989). *Oxford advanced learner's dictionary of current English* (4th ed). Oxford: Oxford univ. press.
Ingwersen, P. (1992). *Information Retrieval Interaction*. London: Taylor Graham Publ.
Ingwersen, P. (1994). The cognitive perspective in information retrieval. *Int. Forum of Information and Documentation, 19*(2), 25–32.
Iyer, H. (1992). Subject representation and entropy. *International Classification, 19*(1), 65–71.
Jachna, J. S., Powsner, S. M., & Miller, P. L. (1993). Augmenting GRATEFUL MED with the UMLS Metathesaurus: an initial evaluation. *Bulletin of Medical Library Association, 81*(1), 20–28.
Jacob, E. K. (1994). Classification and crossdisciplinary communication: breaching the boundaries imposed by classificatory structure. In H. Albrechtsen, S. Oernager (Ed.), *Knowledge organization and quality management: Proc. Third Int. ISKO Conference*, 20–24 June 1994, Copenhagen, Denmark (Advances in knowledge organization, Vol. 4, pp. 101–108). Frankfurt/Maine: Indeks Verlag.
Jamieson, A. J., Dolan, E., & Declerck, L. (1986). Keyword searching vs. authority control in an online catalog. *J. Academic Librarianship, 12*, 277–283.
Jones, S. (1993). A thesaurus data model for an intelligent retrieval system. *J. Info. Sci., 19*, 167–178.
Kim, C. (1982). Retrieval language of social sciences and natural sciences: a statistical investigation. *J. Amer. Soc. Info. Sci., 33*(1), 3–7.
Körner, S. (1970). *Categorial frameworks*. Oxford: Blackwell.
Kristensen, J. (1993). Expanding end-users' query statements for free text searching with a search-aid thesaurus. *Information Processing and Management, 29*, 733–744.
Kuller, A. B., Wessel, C. B., Ginn, D. S., & Martin, T. P. (1993). Quality filtering of the clinical literature by librarians and physicians. *Bulletin of Medical Library Association, 81*(1), 38–43.
Larson, R. R. (1991a). Between Scylla and Charybdis: subject searching in the online catalog.. *Advances in Librarianship, 15*, 175–236.
Larson, R. R. (1991b). Classification clustering, probabilistic information retrieval, and the online catalog. *Library Quarterly, 61*, 133–173.
Lawrence, G. S. (1985). System features for subject access in the online catalog. *Library Resources and Technical Services, 29*(1), 16–33.
Lu, X. (1990). Document retrieval: a structural approach. *Information Processing and Management, 26*, 209–218.
Mandel, C. A., & Herschman, J. (1983). Online subject access-enhancing the library catalog. *Journal of Academic Librarianship, 9*, 148–155.
Mansfield, U. (1982). The systems movement: an overview for information scientists. *J. Amer. Soc. Info. Sci., 33*, 375–382.
Masarie, F. E., & Miller, R. A. (1987). Medical subject headings and medical terminology: an analysis of terminology used in hospital charts. *Bulletin of Medical Library Association, 75*(2), 89–94.
Mattessich, R. (1982). The systems approach: its variety of aspects. *J. Amer. Soc. Info. Sci., 33*, 383–394.
McJunkin, M. C. (1995). Precision and recall in title keyword searches. *Information Technology and Libraries, 14*, 161–171.
Miller, G. I. (1980). *Методологические проблемы предметизации (Сравнительный анализ принципов построения языки предметных рубрик в СССР и США)*. [Methodological problems of subject analysis. (The comparative analysis of the construction principles of subject headings language in USSR and USA)]. Moscow: Nauka.
Miller, U. (1996). The 'Sport Database': Some comments. *Online & CD-ROM Review, 20*(2), 67–74.
Milstead, J. L. (1992). Methodologies for subject analysis in bibliographic databases. *Information Processing and Management, 28*, 407–431.

Wilkes, A. (1995). Subject searching in two online catalogs: authority control vs. non-authority control. *Cataloging and Classification Quarterly*, *20*(4), 57–79.
Willett, P. (1988). Recent trends in hierarchic document clustering: a critical review. *Information Processing and Management*, *24*, 577–597.
Yngve, V. H. (1981). Stoic influences in librarianship: a critique. *J. Library History*, *16*(1), 92–105.

ELSEVIER

International Journal of Medical Informatics 53 (1999) 29–41

International Journal of
Medical
Informatics

Development of a cross-thesaurus with Internet-based refinement supported by UMLS

Stefano Dessena, Angelo Rossi Mori *, Elena Galeazzi

Reparto Informatica Medica, Istituto Tecnologie Biomediche CNR, Roma, Italy

Received 25 February 1998; received in revised form 22 March 1998; accepted 28 March 1998

Abstract

Combinatorial terminological systems are appearing, to solve the issues related to flexibility and precision of representation requested by modern healthcare information systems and in particular by messaging standards. The development of a robust system of descriptors ('cross-thesaurus') is a crucial activity in the production of combinatorial terminological systems. We developed a tool (I-BROWSE) to produce a cross-thesaurus by analysing existing terminological corpora. To facilitate the work of experts and to produce re-usable results, our application interacts via the Internet with the UMLS® Knowledge Sources Server. We applied our tool on 6372 dissections on surgical procedures produced in the project GALEN-IN-USE, as a part of the internal Quality Assurance Program. Support from UMLS seems mostly promising about descriptors on < anatomy >, < pathological structure >, < pathological process > and < physiological process >. Additional assistance can be given to domain experts on less frequent descriptors on pervasive modifiers. We plan to apply our tool also to production of terminological standards in CEN, as a part of a world-wide process of gradual convergence and transformation of coding systems into second-generation systems and terminological services. © 1999 Elsevier Science Ireland Ltd. All rights reserved.

Keywords: Terminology; Standardisation; Thesauri development; UMLS; GALEN; CEN/TC251/WG2

1. Introduction

Users of health care informatics require new terminological systems to support rationalisation of care provision, in particular to achieve more effectiveness, better organisation and to control appropriate use of resources.

Recent dramatic developments in health care terminology, mainly under pressure of messaging standards, are completing a transition from statistics to routine representation

* Corresponding author. Present address. CNR-ITBM, viale Marx 15, 00137 Roma, Italy; e-mail: angelo@color.irmkant.rm.cnr.it

1386-5056/99/$ - see front matter © 1999 Elsevier Science Ireland Ltd. All rights reserved.
PII S1386-5056(98)00092-6

of details, in a new pragmatic perspective (see Table 1), a second generation of terminological systems is coming [15].

In fact, new systems are appearing, based on compositional methods, to achieve the maximum of flexibility and expressiveness [1].

Mapping among existing systems and systematic development of new ones should be based on a pre-defined set of elementary building blocks (from now on called 'descriptors'), organised into a 'cross-thesaurus' derived from the individual systems, but eventually not depending on them [15].

Every phrase of a terminological system can be systematically represented by a 'dissection', i.e. by a combination of descriptors according to a set of composition rules.

A dissection is therefore a precise, unambiguous, structured intentional definition in a well-defined system of concepts.

A crucial activity in this advanced scenario is the production of a robust and comprehensive cross-thesaurus.

It involves many kind of difficult decisions on descriptors, regarding for example: appropriate level of granularity; unambiguous names; and systematic categorisation. These decisions should be based on comparison of similar phrases, understanding of the various shades of meaning, study of words with overlapping meanings.

We developed a software application, called I-BROWSE, to build a cross-thesaurus from the simultaneous analysis of existing terminological corpora, according to an original approach.

To facilitate the work of experts and to produce more re-usable results, our application now interacts via the Internet with the Meta-1 component of the UMLS® Knowledge Sources Server [2].

In fact terminological data from UMLS assist the experts in refining the descriptors and their organisation. Candidate descriptors from terminological corpora can be compared with elementary concepts already present in Meta-1 (e.g. in MeSH and SNOMED axes) and their organisation can be harmonised with known structures (e.g. the semantic network in UMLS) [22].

2. Materials and methods

In this section we introduce the language and the methods about second-generation systems. Then we present our tool and its integration to the UMLS® Knowledge Source Server. Finally we describe the corpus on which we applied our tool.

2.1. Our methodology for co-operative development of second-generation systems

We developed a methodology to produce a second-generation terminological system by systematic analysis of a set of first-generation ones [3], according to the European Prestandard ENV 12264:1997 [4].

Table 1
Towards a second-generation of terminological systems

First generation	Second generation
Coding	Representing
Discard details	Preserve details
Statistics	Assistance to care provision (organisation, planning, control), access to knowledge
Off-line	On-line
Large 'monothematic' coding system	Collection of small value sets and answer lists
Classifications, thesauri, nomenclatures	Multifunctional systems and services
ICD, ICPC, CPT	Read codes, LOINC, ICNP, SDM, Global Medical Devices Nomenclature

The methodology is iterative and evolutionary. It allows for involvement of an increasing number of domain experts across structured cycles.

Let us recall here, in synthesis, the process carried out by the domain experts on each phrase of the corpora.

2.2. The process of terminological analysis

Consider a rubric in a coding system, e.g.:
other heart revascularisation (ICD-9-CM 36.3)
in the context of higher-rank rubrics, e.g.:
operation on vessels of heart (ICD-9-CM 36)

The experts should convey the information from the context into a paraphrase. In principle, they are not allowed to add new details, i.e. they should just make explicit the information understood in the original phrase and fix potential ambiguities, e.g.:

other surgical operation on blood vessels serving heart for vascular perfusion of heart

From this paraphrase they produce a dissection (i.e. a structured representation of a paraphrase in a precise language), e.g.:

MAIN OTHER surgical operation
  ACTS_ON blood vessel
    SERVES heart
  TO_ACHIEVE vascular perfusion
    ACTS_ON heart

A dissection is made of semantic links (e.g. ACTS_ON, SERVES, TO_ACHIEVE) and descriptors (e.g. surgical operation, heart, blood vessel, vascular perfusion).Descriptors belong to categories, e.g.:

<deed>	surgical operation
<process>	vascular perfusion
<anatomy>	blood vessel, heart

A thesaurus can be developed from an individual terminological system to organise a set of descriptors able to systematically represent the phrases of that terminological system. The harmonisation of descriptors (names, granularities, shades of meaning) and categories across various terminological systems produces a cross-thesaurus [15,12].

Dissections obey to a series of patterns (i.e. allowed combinations among categories and links, making the categorical structure to the domain), that can be nested according to predefined rules, e.g.:

1. MAIN <deed>
2. <deed> ACTS_ON <anatomy>
3. <deed> TO_ACHIEVE <process>
4. <process> ACTS_ON <anatomy>
5. <anatomy> SERVES <anatomy>

Reserved keywords provide instructions for further processing of the dissection (e.g. MAIN, OTHER).

In this way the set of dissections can be used as intermediate representation for a further step of semi-automatic translation into a formal system, as in the process adopted in GALEN-IN-USE [5,13,16,18]. In fact, dissections are the compromise between easiness of usage for domain experts and richness of representation and precision for further computer-based processing [18].

2.3. The bottle-neck is a stable cross-thesaurus

Our methodology at first produces simultaneously an approximation of the various components of a second-generation system: cross-thesaurus, categorical structure and, if requested, a set of dissections [15]. These components are then refined the one from the other during various development cycles. The translation into a formal system (third-generation) [5] and the reconstruction from it of phrases in a pseudo-natural language [16] provide important feedback's.

We applied that methodology in different settings:

- In our CEN activities, namely for vital signs [6], medical devices [7,8], headings for patient record [20,21].
- In explorative actions [9] on LOINC [10] and SNOMED-DICOM Microglossary [11].
- For co-operative production of dissections in GALEN-IN-USE [12], where four centres participate in the population of a model of surgical procedures [16].

A bottle neck was to arrange an interim cross-thesaurus to start the process with a reasonable amount of experimental data about most pervasive constructs (descriptors, categories, co-occurrences). This was the main reason for the development of a specific tool.

2.4. Our toolkit: I-BROWSE

I-BROWSE is an application to analyse various collections of phrases (from nomenclatures, classifications, etc.) in order to build rapidly a cross-thesaurus.

Initially each phrase is analysed automatically to extract the occurrences of predefined lexemes (words, or idioms and parts of words) and of unknown words. The lexicon of words, idioms and parts of words is registered in evolving tables, updated during this initial phase. After an automatic step, manual revisions of the tables and the related parsing can be carried out.

New candidate descriptors are produced semi-automatically from lexemes and can be grouped by the domain expert in categories and subcategories, redefined or adapted—when necessary—for the particular corpus.

Lexemes preserve the strings as in the original corpus, descriptors use arbitrary strings with the intent of conveying to developers an unambiguous meaning.

Descriptors, lexemes, categories and phrases are stored into a relational database (Microsoft ACCESS). The tool allows to browse all of the above mentioned types of entities in relation to any given type of entity, with the primary goal of refining number and names of descriptors and their grouping into categories. Descriptors and lexemes can be ordered and filtered by frequency, in order to allow the expert to focus on the most pervasive items first.

An early prototype (using FoxBASE) was previously used to build the cross-thesaurus for the European Prestandard ENV 12611:1996 on medical devices [7]. The tool is being extended to process dissections produced by the 4 centres of GALEN-IN-USE and to carry out Quality Assurance on this material, producing statistics on usage of descriptors, links and categories, looking at co-occurrences, making reports on potential issues (e.g. recognise different usages by centres, provide examples for controversial problems).

In our first experiences, domain experts were completely free to define their own descriptors and categories. We were not willing to impose an overload (due to some compliance to external constraints) not balanced by benefits, e.g. from re-use of work from other sources. Nevertheless, it was clear since the beginning that freedom of experts had to be partially limited, to allow re-use of descriptors and categories in different domains.

A mechanism to maximise the return to domain experts with a low burden was therefore studied, involving an integration to UMLS [22].

2.5. The services of UMLS on the web

UMLS recently made available a server to access the Knowledge Sources via the Internet [2]. Access to the system is provided

through a command-line interface, an Application Programming Interface (API) and through the World Wide Web. The API consists of simple functions for establishing and breaking a connection to the server, sending queries and receiving results from the server.

An additional feature was therefore introduced into the environment of I-BROWSE, that allows to query the Metathesaurus® database and to integrate the results into the structure of our existing relational database. Our application is now able to retrieve automatically the suitable information from the Metathesaurus, during predefined phases of the process [22].

During the construction of the tables on lexemes, our systems asks, for each lexeme found during the parsing phase, if there is an exact match with one of the variants of a Meta-1 term as a whole and in case of success it receives:

- The unique identifier for concept in Meta-1 (CUI).
- The preferred term.
- The semantic types.
- The code in SNOMED International [14].

Moreover, our application allows the experts to retrieve on request more specific additional data, in particular:

- The Meta-1 terms where our lexeme (or a part of it) appears as a substring.
- Any other information via an ad hoc interface (but results of the query remain independent from the other tables of our application).

The goal is to provide hints to a domain expert in order to facilitate the refinement of more re-usable descriptors, recognise potential conflicts and ambiguities, assist in the shaping of the peculiar categories and subcategories for a cross-thesaurus and in the assignment of descriptors to categories.

2.6. The corpora used for the experiment

Our test set was taken from the current work on surgical procedures in the European Project GALEN-IN-USE [13]. Its goal is a formal model with significant coverage of the domain.

Rubrics come from four coding systems: ICD-9-CM, Nordic Classification of Surgical Procedures, Dutch adaptation of ICPM and NCAM, a system being developed in France [16]. At the moment of the analysis presented in this paper (January 1998), 6372 dissections have been produced by four centres (in Italy, Sweden, the Netherlands and France), debugged according to formal and semantic rules, successfully translated in the GRAIL formalism and loaded into the 'Common Reference Model'. The model will be revised and completed by the end of 1998.

As a part of the internal Quality Assurance Program in GALEN-IN-USE, the team at CNR is analysing the corpus of dissections, to suggest options for improvement of the model. One of the tasks is to revise the set of descriptors (now that a significant amount of experimental material is available, in a format suitable for processing), to make them more re-usable for next cycles and for other applications.

A preliminary statistical analysis was carried out on descriptors and on the successes in the linkage to Meta-1.

Moreover, many descriptors in GALEN-IN-USE (e.g. 'artery of the lower extremity') are more complex than the usual ones in I-BROWSE. In fact centres use combinations of elementary descriptors to avoid too long dissections. Therefore we also carried out two different kinds of analysis:

- The first is general, about the correspondence of our descriptors into the Metathesaurus.

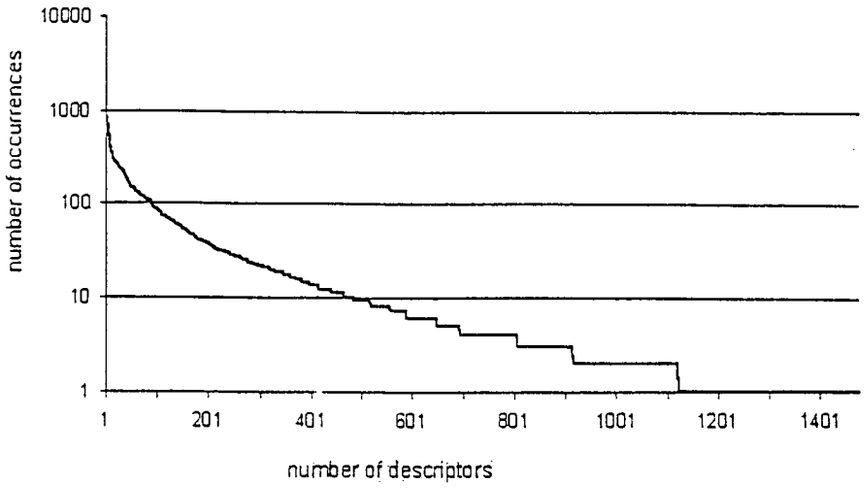


Fig. 1. Distribution of the 1476 descriptors by number of occurrences (log scale).

- The second is limited descriptors made of a single word, that are good candidates for highly re-usable units.

3. Results

3.1. Basic statistics on our corpora

Until January 1998, the activity of GALEN-IN-USE which has been developing the model on surgical procedures has produced 6372 dissections. It used 1476 descriptors, for a total of 36262 occurrences, with an average respectively of 5.7 occurrences/dissection and 24.6 occurrences/descriptor.

In other words, each dissection was made of an average of 5.7 lines (the first line contains the keyword 'MAIN' followed by a descriptor, each subsequent line contains a semantic link and a descriptor; see Section 2.2).

Each descriptor has been used until now 24.6 times in average. The maximum number of occurrences was for 'excising' (1120 times) and 'removing' (849 times). On the other side, 356 descriptors (24%) were used only once and 207 (14%) are used twice. The complete distribution is presented in Fig. 1.

The cumulative distribution of the occurrences of descriptors is presented in Fig. 2. On the x-axis there is the rank order of the descriptors by frequency (no. 1 is the most frequent, no. 1476 is the less frequent). On the y-axis there is the cumulative distribution (expressed in percent) from rank 1 to x.

Relevant points for discussion are:

- 5% of descriptors (i.e. the most used 74 descriptors) account for 54.5% of occurrences.
- 20% of descriptors (i.e. the most used 296 descriptors) account for 84.2% of occurrences.

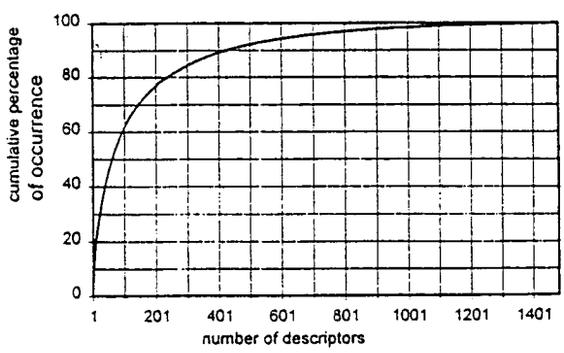


Fig. 2. Cumulative percentage of 35715 occurrences of 1476 descriptors in the sample of 6372 dissections.

Table 2
Examples of descriptors by category

Category	Examples
Human anatomy	Face, sacral vertebra, abdomen
Surgical deed	Excising, destroying, dividing
Device	Catheter, pacemaker, laparoscope
Pathological structure	Cyst, aneurysm, hernia, polyp
Approach	Transurethral, percutaneous, closed
Extent value	Partial, local, extended, total
Material	Thrombolytic agent, air, fluid
Pathological process	Infarction, luxation, trauma
Side	Left, unilateral, bilateral
Position value	Distal, lateral, upper, anterior
Temporal value	Permanent, recent, postoperative
Health care activity	Cure, investigation, prevention
Physiological process	Vascular perfusion, respiration
Other features	Benign, billroth2 method, rigid

- 30% of descriptors (i.e. the most used 443 descriptors) account for 90.6% of occurrences.

During the production of the dissections, the four centres assign each new descriptor to one category. This category is selected from an evolving set of detailed semantic types and subtypes, organised in a hierarchy, at present the set of types and subtypes for surgical procedures is made of 52 categories, in a tree with up to four hierarchical levels.

This assignment is then revised by the University of Manchester, that is mapping the descriptors to GRAIL concepts [5]. From the 1476 descriptors of the sample, 1323 descriptors were already revised in this way. The following statistics on categories are therefore based on this revised subset.

For the purposes of this paper, we use here a simplified set of 14 categories, as presented in Table 2.

Detailed distribution of revised descriptors and of their occurrences by category are presented in Table 3. For each category, we also calculated the average amount of occurrences per dissections and the total occurrences for each descriptor in our current sample.

3.2. Success of retrieval of concepts from Meta-1

Descriptors are internally denominated by unambiguous artificial strings. Each of them corresponds to one or more strings (lexemes), that actually appear in the corpora.

The 1325 descriptors corresponded in our corpus to 1689 different strings (1.28 different strings for each descriptor).

We searched for an exact match of our 1689 lexemes in the Meta-1 section of UMLS. For each concept in Meta-1, we asked also for its semantic type(s).

The success rate was 40.0% within the whole Meta-1 collection and 31.0% within the SNOMED phrases. A detailed table by category is presented in Table 4.

We also asked for partial matching of our lexemes as substrings within phrases of Meta-1. The success rate raised to 61.3%.

Among our lexemes, 764 were made of a single word (45.2% of the strings). If we focus on these strings, the success rate was respectively 49.7% for Meta 1, 35.1% for SNOMED and 75.1% for partial matching of our lexemes as substrings contained in Meta-1 phrases. Results are presented in Table 5.

Some of our strings (97 strings, 14.3%) matched multiple meanings (homographs), or were corresponding to multiple semantic types. For example, 'removal', 'excision' and 'resection' could be considered as < research activity > or as < therapeutic or preventive procedure > . Details are presented in Table 6.

Table 3
Basic statistics about descriptors in our corpus

	Different descriptors		Total occurrences		Occurr/dissect	Occurr/descr
	No.	%	No.	%		
Human anatomy	622	47.0	13 746	38.5	2.16	22.1
Surgical deed	166	12.5	11 380	31.9	1.79	68.6
device	131	9.9	2767	7.7	0.43	21.1
Pathological structure	90	6.8	2369	6.6	0.37	26.3
Approach	58	4.4	1593	4.5	0.25	27.5
Extent value	9	0.7	817	2.3	0.13	90.8
Material	48	3.6	634	1.8	0.10	13.2
Pathological process	47	3.6	624	1.7	0.10	13.3
Side	6	0.5	433	1.2	0.07	72.2
Position value	10	0.8	278	0.8	0.04	27.8
Temporal value	11	0.8	258	0.7	0.04	23.5
Health care activity	11	0.8	215	0.6	0.03	19.5
Physiological process	9	0.7	61	0.2	0.01	6.8
Other features	105	7.9	540	1.5	0.08	5.1
Total	1323		35 715		5.60	27.0

Rows are ordered by number of occurrences.

Finally, 14 couples and one quadruple of descriptors mapped to the same concept in Meta-1 (i.e. they were considered as synonyms in Meta-1, having the same CUI). For example, the four centres were producing different descriptors for: 'puncture' and 'drilling' (CUI 34117); 'cranium' and 'skull' (CUI 37303); 'chest' and 'thorax' (CUI 39992); the quadruple is 'removal', 'excision', 'ablation' and 'resection' (CUI 15252).

4. Discussion

Building a thesaurus is a complex task. The development of a cross-thesaurus, to comply with the needs of complete, systematic representation of multiple terminological systems, is still more complex.

Even if driven by general ontological principles, the production and refinement of a cross-thesaurus is mainly a bottom-up experimental activity.

The harmonisation of cross-thesauri from different healthcare domains is our long-term goal, that should pass before by 'local' harmonisation of the descriptors behind a set of homogeneous terminological systems.

We can assist the experts in this work, but they are not accepting predefined solutions as such, just because most thesauri developed for other purposes are not easily re-usable for this new peculiar task, the systematic representation of phrases by structured definitions from a precise, large system of concepts.

Most of our descriptors are elementary concepts of a granularity similar to the one of a thesaurus (e.g. MeSH). Less frequent are complex descriptors, i.e. pre-co-ordinate forms on frequent combinations of more elementary descriptors.

We use descriptors to represent details in source phrases and not for indexing. There-

Table 4
Success rate (absolute and percent) in exact matching of our lexemes to Meta-1 concepts and SNOMED codes

Strings	All	Found in Meta-1		Found in SNOMED		Appears as substring	
Categories		No.	%	No.	%	No.	%
Human anatomy	774	416	53.7	373	48.2	554	71.6
Surgical deed	245	38	15.5	26	10.6	112	45.7
Device	151	36	23.8	5	3.3	72	47.7
Pathological structure	111	66	59.5	32	28.8	74	66.7
Approach	93	12	12.9	12	12.9	31	33.3
Extent value	12	4	33.3	4	33.3	9	75.0
Material	58	31	53.4	20	34.5	44	75.9
Pathological process	50	23	46.0	11	22.0	33	66.0
Side	12	4	33.3	3	25.0	6	50.0
Position value	11	10	90.9	9	81.8	11	100.0
Temporal value	18	5	27.8	3	16.7	14	77.8
Health care activity	15	5	33.3	3	20.0	8	53.3
Physiological process	13	7	53.8	7	53.8	10	76.9
Other features	126	18	14.3	16	12.7	58	46.0
Total	1689	675		524		1036	
Average			40.0		31.0		61.3

The last two columns refer to the success of partial matching of our lexeme as a substring in the Meta-1 phrases.

fore our set of descriptors includes the ones of an usual thesaurus, but also many additional concepts not relevant for current information retrieval tasks.

Our expectation was therefore to see various classes of descriptors according to their behaviour in our queries:

- Common 'monoaxial' concepts as the ones in MeSH (for indexing) and in Topography, Morphology, Function, Living organisms of SNOMED (for 'primary' compositions).
- Generic concepts on procedures, diseases, devices (typically present in Meta-1 within more complex terms, but also as isolated concepts).
- Common modifiers, scarcely present in isolation, but normally part of Meta-1 terms.
- Peculiar modifiers, used in highly specialised (local) nomenclatures, but not present in Meta-1 terms.

We show below how our results confirm this hypothesis.

4.1. Characterisation of descriptors in the corpus

Our results are based on the analysis of about one half of the whole field of surgical procedures. We were expecting that most of the generic < deeds > were covered, but of course the descriptors of < anatomy > will perhaps add another 50% in the next months.

From the figures of the 'occurrences by descriptor' and 'occurrences in a dissection' in Table 3 we clearly see that

- The average number of occurrences for a (revised) descriptor is about 27, corresponding to 5.6 descriptors per dissection.
- descriptors of < deed > have an average of nearly 70 occurrences (there is a subset of generic descriptors, plus a large number of speciality-related ones); they appear

Table 5
Success rate(absolute and percent) in exact matching of our single- word lexemes to Meta-1 concepts and SNOMED codes

Strings	All	Found in Meta-1		Found in SNOMED		Appears as substring	
Categories		No.	%	No.	%	No.	%
Human anatomy	264	187	70.8	162	61.4	229	86.7
Surgical deed	194	37	19.1	25	12.9	107	55.2
Device	60	27	45.0	2	3.3	37	61.7
Pathological structure	63	49	77.8	21	33.3	54	85.7
Approach	13	1	7.7	1	7.7	9	69.2
Extent value	9	4	44.4	4	44.4	9	100.0
Material	29	18	62.1	12	41.4	23	79.3
Pathological process	15	12	80.0	4	26.7	15	100.0
Side	7	4	57.1	3	42.9	5	71.4
Position value	11	10	90.9	9	81.8	11	100.0
Temporal value	15	5	33.3	3	20.0	12	80.0
Health care activity	7	4	57.1	2	28.6	6	85.7
Physiological process	7	4	57.1	4	57.1	6	85.7
Other features	70	18	25.7	16	22.9	51	72.9
Total	764	380		268		574	
Average			49.7		35.1		75.1

The last two columns refer to the success of partial matching of our lexeme as a substring in the Meta-1 phrases.

1.79 times per dissection, once as MAIN deed and the rest as additional procedures or as techniques (see discussion in [17]).

- A few descriptors of < extent > and < side > have more than the double of the average (respectively 90 and 70 occurrences), but they appear respectively only on 13% and 7% of dissections.
- Descriptors of < device >, < pathological structure >, < approach >, < material >, < pathological process > are largely used across the whole field.

Table 6
Number of choices explicitly suggested to the domain expert by the results of Meta-1 queries

Choices	Meta-1 concepts	%
1	578	85.6
2	88	13.0
3	9	1.3
Total	675	100

- Descriptors of < anatomy > are present more than twice per dissection. A large number of anatomic descriptors is generic, but many different descriptors are involved in the different chapters, so that they account for 50% of the total amount of descriptors.

Descriptors follow the '20/80' rule about efficiency: 20% of the resources (at present, 296 descriptors) account for 84% of the results (at present, 30545 occurrences). They can be carefully revised and included in the planned revision of the CEN pre-standard ENV 1828 on the categorical structure for surgical procedures [23].

Among that 20% of descriptors, we can foresee that about 100 descriptors (at present, 74) will be very pervasive and will cover about one half of the occurrences (at present, 54.5%). We can therefore plan additional studies (e.g. [19]) on large samples of phrases, to discover all their potential shades of meaning.

4.2. Linkage to Meta-1 and SNOMED

Our goal for linking I-BROWSE to Meta-1 and SNOMED was to assist our experts with reference material. From Tables 4 and 5, we see that the potential for assistance from UMLS to experts about < anatomy > is high, the amount of hits in Meta-1 is large (54%) as well as the amount of SNOMED codes (48%). We expect that after a careful revision and a purposive harmonisation those figures can reach even higher values.

In general, for descriptors that are common in our corpus we can expect that our tool will speed the work of categorisation and facilitate discovering of potential homographs and synonyms. On < deed >, we were aware that descriptors required ad hoc resources, thus specific results on that topic were achieved by other approaches [17].

We have also a high number of hits—more than 50%—for descriptors that are less common in our corpus, but pervasive in healthcare (such as < material >, < pathological structure >, < pathological process > or < physiological process >).

Here our corpus does not provide enough data for a stable model, but we can increase robustness and universality of these descriptors by the interaction with the UMLS Sources.

We expect two other modalities for additional assistance from UMLS, on discovering ambiguities in our descriptors:

- Explicit choices proposed by the tools (see Table 6) for homographs and concepts with multiple semantic types.
- Browsing of Meta-1 terms containing the descriptor as a substring.

In this latter case, the level of hits is extremely high, 61.3% for the whole set of our descriptors, 75.1% for the descriptors corresponding to single-word lexemes. In some small categories we have 100% of hits and in anatomy (264 descriptors) we reached a level of 86.7% of matches for single-word lexemes. In other words, our experts will be able to reason, in most cases, on a wider set of phrases, to achieve more robust decisions on the modelling of descriptors.

At the moment we allow the expert to download up to 50 Meta-1 terms per descriptor and we store them into our database, for further systematic comparisons and thinking (see for example the comparative study on 'removal' and 'excision' carried out on the dissections and on the terms extracted from Meta-1 [19]).

4.3. Benefits from the linkage to UMLS

In synthesis, the linkage to UMLS allows to:

- Refer to standard concepts (in particular SNOMED), looking also at preferred terms and variants.
- Extend the number of observed occurrences to refine the meaning of a descriptor, by browsing Meta-1 terms containing a given substring.
- Facilitate categorisation (even in the case when the expert wants a different organisation of the semantic types).

The close link to the UMLS Knowledge Sources allows also to provide, in future, terminological data to the various Committees that maintain these Sources (in particular SNOMED) in order to achieve a gradual convergence.

5. Conclusions

The linkage to UMLS produces the expected results. The domain expert can be supported in the task of building a robust cross-thesaurus by the knowledge provided by queries to Meta-1. We can also envisage a

feedback to the Committees responsible for the maintenance of the Knowledge Sources embedded in UMLS (in particular SNOMED International), in order to facilitate gradual convergence among systems.

The tool seems to realise a suitable compromise, i.e. to provide enough benefit to the experts to facilitate convergence (by reference to authoritative Sources and purposive, systematic rearrangements of terminological data), without overloading them with too strong constraints of coherence.

Generalisation from existing systems in a given domain is already a very difficult task and we cannot ask the experts also to worry for universality of descriptors across different domains. By the use of our tool, linked to UMLS, we hope that the browsing capabilities and the potential for simple comparisons will allow the experts to produce a first set of reasonably robust and coherent cross-thesauri (each satisfying a particular domain), that in future could converge into the same 'universal' system with a reduced effort.

We can optimise our efforts, by concentrating deep co-operative studies on one hundred of descriptors that account for half of the occurrences and by revising carefully the systematisation of 300 to 500 descriptors that will account for 80 to 90% of the occurrences.

Healthcare messages will contain the original phrases (and codes) and a number of structured details for their correct interpretation by computer. With the approach described in this paper, we will be able—very soon—to convey in a systematic way at least the most pervasive details in the most significant domains of healthcare.

Acknowledgements

Work partially performed within the European project GALEN-IN-USE (HC 1018).

References

[1] A. Rossi Mori, F. Consorti, E. Galeazzi, Standards to support development of terminological systems for healthcare telematics, Method Inf. Med. 1997 (in press)

[2] http://www.nlm.nih.gov/pubs/factsheets/umlskss.html

[3] E. Galeazzi, A. Rossi Mori, F. Consorti, A. Errera, A methodology to build conceptual models in medicine, in: C Pappas, et al. (Eds.), Medical Informatics Europe 97, IOS Press, Amsterdam, 1997, pp. 280–284.

[4] CEN ENV 12264: 1995. Medical Informatics—Categorical structure of systems of concepts—Model for representation of semantics.

[5] J.R. Rogers, W.D. Solomon, A.L. Rector, P. Pole, P. Zanstra, E. van der Haring, Rubrics to dissections to GRAIL to classifications, in: C. Pappas, et al. (Eds.), Medical Informatics Europe 97, IOS Press, Amsterdam, 1997, pp. 241–245.

[6] CEN/TC251 PT5-021 Vital Sign Information Representation, 1996.

[7] CEN ENV 12611:1996 Medical Informatics-Categorical structure of systems of concepts on medical devices.

[8] L. Rystrøm, A. Rossi Mori, Development of a standard for the organisation of medical device nomenclatures, in: J. Brender, et al. (Eds.), Medical Informatics Europe 96, IOS Press, Amsterdam, 1996, pp. 259–263.

[9] A. Rossi Mori, E. Galeazzi, F. Consorti, W.D. Bidgood Jr., Conceptual Schemata for Terminology: A Continuum from Headings to Values in Patient Records and Messages. JAMIA Supplement, Proceedings of the Twenty-First Annual Meeting of the American Medical Informatics Association (in press).

[10] http://www.mcis.duke.edu:80/standards/termcode/loinc.htm

[11] W.D. Bidgood Jr. (Ed.), SNOMED DICOM Microglossary. Coll. Am. Path. Northfield, IL 1997. http://www.snomed.org/sdm/sdm.htm

[12] E. Galeazzi, A. Rossi Mori, F. Consorti, A. Errera, MEDITA: the reference methodology for co-operative production of dissections by terminological analysis. GIU Report MET01/97, available through ref. 13.

[13] GALEN and GALEN-IN-USE documentation (1992–96): http://www.cs.man.ac.uk/mig/galen

[14] R.A. Coté, D.J. Rothwell, J.L. Palotay, R.S. Beckett, L. Brochu (Eds.), The Systematised Nomencla-

ture of Human and Veterinary Medicine. Northfield, IL. College of American Pathologists 1993.

[15] A. Rossi Mori, F. Consorti, E. Galeazzi, P. Merialdo, A second generation of terminological systems is coming, in: C. Pappas, et al. (Eds.), Medical Informatics Europe 97, IOS Press, Amsterdam, 1997, pp. 436–440.

[16] J.M. Rodrigues, B. Trombert-Paviot, R. Baud, J. Wagner, P. Rusch, F. Meusnier, GALEN-IN-USE an EU Project applied to the development of a new National coding system for surgical procedures, NCAM, in: C. Pappas, et al. (Eds.), Medical Informatics Europe 97, IOS Press, Amsterdam, 1997, pp. 897–901.

[17] A. Rossi Mori, A. Gangemi, G. Steve, F. Consorti, E. Galeazzi, An ontological analysis of surgical deeds, in: E. Keravnou, et al. (Eds.), Artificial Intelligence in Medicine 97, Springer, Berlin, 1997, pp. 361–372.

[18] A.L. Rector, A. Rossi Mori, F. Consorti, P. Zanstra, Practical development of re-usable terminologies, Int. J. Med. Inf. (in press).

[19] F. Consorti, A. Rossi Mori, E. Galeazzi, A comparative study of the meanings of 'removal' and 'excision'. Report ITBM-CNR for GALEN-IN-USE, December 1997.

[20] A. Rossi Mori, F. Consorti, Exploiting the terminological approach from CEN/TC251 and GALEN to support semantic interoperability of healthcare record systems, Int. J. Med. Inf. (in press).

[21] A. Rossi Mori, F. Consorti, E. Galeazzi, Preliminary material for the Project Team 'Domain term list' (WI 1.8, part 2), presented at the Joint Working Group Meeting of CEN/TC 251 Healthcare Informatics, Seville, January 1998.

[22] S. Dessena, Estensione di basi di conoscenza terminologiche via Internet. Tesi di Laurea in Ingegneria Informatica. Roma, Ottobre 1997 (in Italian).

[23] CEN ENV 1828:1995. Health care informatics—Structure for classification and coding of surgical procedures. Brussels: CEN, 1995.

ELSEVIER

Neurocomputing 48 (2002) 819–846

NEUROCOMPUTING

www.elsevier.com/locate/neucom

Hierarchical word clustering — automatic thesaurus generation

Victoria J. Hodge[1], Jim Austin*

Department of Computer Science, University of York, York YO10 5DD, UK

Received 23 December 1999; accepted 13 August 2001

Abstract

In this paper, we propose a hierarchical, lexical clustering neural network algorithm that automatically generates a thesaurus (synonym abstraction) using purely stochastic information derived from unstructured text corpora and requiring no prior word classifications. The lexical hierarchy overcomes the *Vocabulary Problem* by accommodating paraphrasing through using synonym clusters and overcomes *Information Overload* by focusing search within cohesive clusters. We describe existing word categorisation methodologies, identifying their respective strengths and weaknesses and evaluate our proposed approach against an existing neural approach using a benchmark statistical approach and a human generated thesaurus for comparison. We also evaluate our word context vector generation methodology against two similar approaches to investigate the effect of word vector dimensionality and the effect of the number of words in the context window on the quality of word clusters produced. We demonstrate the effectiveness of our approach and its superiority to existing techniques. © 2002 Elsevier Science B.V. All rights reserved.

Keywords: Neural network; Hierarchical thesaurus; Lexical; Synonym clustering

1. Introduction

Due to the proliferation of information in databases and on the Internet, users are overwhelmed producing *Information Overload*. Web Search engines can return millions of potential matches to user queries even when more complex (and user-unfriendly) Boolean logic is employed. Web Search engines can be slow,

* Corresponding author. Tel.: +44-1904-432734; fax: +44-1904-432767.
E-mail addresses: vicky@cs.york.ac.uk (V.J. Hodge), austin@cs.york.ac.uk (J. Austin).
[1] This research was supported by an EPSRC studentship.

0925-2312/02/$ - see front matter © 2002 Elsevier Science B.V. All rights reserved.
PII: S0925-2312(01)00675-0

although faster Search Engines are being developed, and matching is often poor (quantity does not necessarily indicate quality) as Search Engines often employ simple keyword pattern matching that takes no account of relevance. Search Engines often simply return the document with the greatest number of keyword occurrences. A methodology to process documents unsupervised, handle paraphrasing of documents, to focus retrieval by minimising the search space and to automatically calculate the document similarity from statistics available in the text corpus is desired. Document may be clustered according to the user's requirements (clustered 'on the fly') and then employ category-specific finer-grained matching techniques.

Word categorisation (encompassing both unsupervised clustering and supervised classification) enables the words to be associated or grouped according to their meaning to produce a thesaurus. In this paper, we focus solely on word clustering as this approach is unsupervised. Clustering does not require pre-generated human classifications to train the algorithm and is therefore less subjective and more automated as it learns from text corpus knowledge only. Word clustering can also overcome the *Vocabulary Problem* cited by Chen et al. [2]. They posit that through the diversity of expertise and background of authors and the polysemy of language, there are many ways to describe the same concept; there are many synonyms. In fact, Stetina et al. [20] postulate that polysemous words occur most frequently in text corpora even though most words in a dictionary are monosemous. Humans are able to intuitively cluster documents from imputed similarity. They overcome the differing vocabularies of authors and the inherent synonymy and polysemy of language. A computerised system must be able to match this ability. For computerised document similarity calculation, an underlying hierarchical synonym clustering is required to enable differing vocabularies to be accommodated. The distances in the hierarchy may be used for word similarity estimation and to score document similarity, thus allowing paraphrased documents to be awarded high similarity scores as their contained words fall into identical or neighbouring synonym clusters. Human generated thesauri are too general; they encompass all senses of words even though many are redundant for a particular domain. They are expensive with respect to construction time, particularly, if a single human knowledge engineer generates the hierarchy. If multiple experts are consulted, then it is very difficult to obtain a single unified hierarchy. Human thesauri also omit certain senses and subdivide others where there is little distinction; they are rather subjective. Automatic methods can be trained generally or domain specifically as required. The hierarchy allows us to focus searching to cohesive clusters therefore minimising the search space for each query. In this paper, we analyse current word categorisation approaches and describe and evaluate our method with respect to the current implementations. We compare our TreeGCS clustering method [7,6] and Sections 3.2 and 3.3 to the self-organising map (SOM) [11] method and then compare three methods for context vector generation where the vector dimensionality and the number of words in the context are varied. We demonstrate the necessity of using high-dimensional vectors to represent the individual words in the documents. High-dimensional vectors ensure that the word vectors are approximately orthogonal and there are no

implicit word dependencies or relationships in the vectors representing the individual words. Therefore, all dependencies and relationships are imputed purely from the relationships between the document words. We demonstrate the superiority of a wider context window when generating the context vectors, illustrating the superior quality clusters and higher stability of the cluster topology produced. Finally, we establish the higher quality of the clusters produced by TreeGCS compared to SOMs. The clusters produced from TreeGCS are similar to the clusters extracted from a benchmark human generated thesaurus.

Our approach is entirely automated and uses only unstructured text corpora as data. The motivation for our approach derives from the patterns present in text. These patterns produce statistical correlations in the context patterns of individual words. We can thus infer the similarities of words from their contexts, as similar words (synonyms) will have similar contexts due to their correlations. Through unsupervised text processing and clustering, we represent semantic relationships by categorising the word co-occurrence patterns. We do not need to generate any linguistic structures, which are complex to produce and tend to be susceptible to linguistic ambiguities. We automatically infer a domain-specific or generalised hierarchical thesaurus as required. We, therefore, surmount the *Vocabulary Problem* [2] by permitting synonym retrieval to match paraphrased documents. We can use the thesaurus to award scores to synonyms using the intra-cluster distances and the inter-cluster distances in the hierarchy.

2. Current methods

Current approaches for textual analysis are multifarious and diverse. The motivations encompass word sense disambiguation, synonym inferencing and both classification and clustering. They include (the following list is not exhaustive but is intended to be broad): contextual methods, WordNet hierarchy methods, clustering methods and SOM methods.

2.1. Contextual methods

These approaches utilise the local neighbourhood of words in a document (the context) to establish lexical similarity and impute synonym groups or disambiguate polysemic words. Yarowsky [22] employs two phases: an iterative bootstrapping procedure and an unsupervised categorisation phase. All instances of a polysemous word are identified in the text corpus. A number of representative samples are selected from each sense set and used to train a supervised classification algorithm. The remainder of the sense sets are trained into the supervised classifier. The classifier may additionally be augmented with one sense per discourse information, i.e. document topic. The classifier can then be used in an unsupervised mode to categorise new exemplars. Stetina et al. [20] postulate that one sense per discourse holds for nouns but evidence is much weaker for verbs. The approach therefore is only suitable for nouns and requires an appraisal of the text corpus

before processing commences to identify the nouns. The method is only partially unsupervised requiring a supervised initial training method; i.e. human intervention which can be time-consuming.

The motivation for Schütze and Pederson [19] is a lexical hierarchy exploiting contextual statistics and requiring no prior data knowledge. The algorithm collects a symmetric, term-by-term matrix recording the number of times that words i and j co-occur in a symmetric window centred about word i in the text corpus, where i and j are any random word indices from the list of all corpus words. Singular-valued decomposition (SVD) is used to reduce the dimensionality of the matrix to produce a dense vector for each item that characterises its co-occurrence neighbourhood. The dense co-occurrence vectors are clustered using an agglomerative clustering algorithm to generate a lexical hierarchy. The method groups words according to their similarity unsupervised rather than some pre-computed thesaurus categories. However, vector dimensionality reduction introduces computational complexity and may cause information loss as the vectors induced represent the meta-concepts and not individual words. Shannon's theory states that *the more infrequent a word, the more information it conveys.* These may well be discarded by SVD. The method does not account for the proximity of the word co-occurrences (co-occurrence is considered from a purely binary perspective). There is no weighting of the co-occurrence according to the two terms' proximity in the context window.

2.2. *WordNet hierarchy*

These methods utilise the human generated hierarchical categorisation of synonyms, hyponyms (IS-A) and metonyms (PART-OF) of WordNet to estimate word similarity and the most appropriate word sense (WordNet lists all senses of words with the most frequently occurring listed first). Li et al.'s [14] method utilises the WordNet synonym, hyponym and metonym hierarchy to assign word similarity according to the distance in the hierarchy. Similarity is inversely proportional to distance. However, the distance of taxonomic links is variable, due to certain sub-taxonomies being much denser than others. Again, the technique relies on an underlying pre-determined word hierarchy and can only process words present in the hierarchy; it could not extrapolate similarities to new words. Human generated thesauri are subjective and rely on sense categorisation decisions made by the human constructor.

2.3. *Clustering*

An unsupervised clustering algorithm derives the word clusters and models of association directly from distributional data rather than pre-determined classes as in Yarowsky. Pereira et al. [17] employ a divisive clustering algorithm for probability distributions to group words according to their participation in particular grammatical relations with other words. In the paper, nouns are classified according to their distribution as direct objects of verbs with cluster membership defined

by $p(c|w)$ (the probability a word belongs to a cluster) for each word rather than hard Boolean classification. Deterministic annealing finds the sets of clusters by starting with a single holistic cluster and increasing the annealing parameter (see paper [17]). As the annealing parameter increases, the clusters split producing a hierarchical data clustering. The approach is limited to specific grammatical relations, requiring a pre-processor to parse the corpus and tag the part-of-speech. At the time of writing, the authors felt the technique required further evaluation.

2.4. Self-organising map (SOM) methods

Word vectors or document vectors form the input vector space of the SOM [11] to permit topological mapping, to infer similarity and categorise words and documents. The aim of Lowe [15] is a topological mapping of contextual similarity exploiting contextual information to derive semantic relationships. Each word in a 29-word vocabulary is associated with a 58-element co-occurrence vector. The value of the nth attribute in the co-occurrence vector reflects the number of times the nth word of the vocabulary has preceded and the $(n+29)$th attribute represents the number of times the nth word has succeeded the keyword where $1 \leqslant n \leqslant 29$. The 58 element vectors form the input vectors for a SOM network. The SOM is labelled by determining the best matching unit for each input vector. The word contexts (labels) are arranged topologically according to lexical and semantic similarity by the SOM. However, the method is inherently susceptible to the scalability problem; vector length grows linearly in relation to lexical size and thus the method is not feasible for a large vocabulary.

Ritter and Kohonen's [18] approach provides the motivation for our system. A topological map of semantic relationship among words is developed on a self-organising feature map. In the initial implementation, each word has a unique, seven-dimensional, unit length vector assigned. The input vector space is formed from the average context in which each word occurs in the text corpus. Semantic similarity is induced from context statistics, i.e., word neighbourhoods using a window of size three, one word either side of the target word, (only nouns verbs and adverbs are used in the method). The method has been extended to WEBSOM [9,8,12] that categorises over one million documents using a window size of three and 90-dimensional word vectors. The approach is entirely unsupervised requiring no human intervention and parallelisable enabling computational speedup. However, SOMs cannot form discrete (disconnected) clusters thus inhibiting the data representation. The clusters have to be determined after the algorithm terminates by hand and this introduces the innate subjectivity of human judgements. Also, the word topography in WEBSOM is single-layered compared to the hierarchical topology we induce using TreeGCS [7,6] and described later in the paper.

2.5. Summary

Many of the methods exposited use Zipf's law [23] and stop-word elimination to reduce the vocabulary of the text corpus to be processed, some even implement

word stemming. Zipf's law implies that a significant portion of the words in a corpus constitutes the words that appear most infrequently whereas frequently occurring words comprise a relatively small portion of the corpus. Many approaches eliminate these infrequent words to decrease vector dimensionality and computational requirements. The designers of these approaches deem that such words provide little discriminatory power for document similarity assessment. We feel that this may discard essential information. Although we do not generate context averages for frequent words, e.g., {the, and, but, etc.}, we include these words in the context averaging of the keywords. For this reason, we use a context size of seven (three words either side of the target word). We demonstrate in Sections 4 and 6 the qualitative improvement of word clustering against a human generated thesaurus and Euclidean distance-based vector approach of a size seven window compared to size three. Ritter and Kohonen [18] and their extrapolations [9,8,12], fix the context window at three and thus have to discard frequent terms, infrequent terms and punctuation, etc. We feel these provide much information and are certainly employed by a human reader when parsing text. Dagan et al. [3] empirically demonstrated that singleton words (words occurring once) were important for parsing concurred by Shannon's theory. An infrequent word occurring only once in two documents may be the key to identifying those documents and should not be discarded from the indexing. The larger window allows us to maximise the lexical information used and minimise the amount of pre-processing required.

3. Our methodology

We cluster words into a synonym hierarchy using the TreeGCS hierarchical clustering neural network that we have developed, described in [7,6] and Sections 3.2 and 3.3. TreeGCS is an unsupervised growing, self-organising hierarchy of nodes able to form discrete clusters. Similar high-dimensional inputs are mapped onto a two-dimensional hierarchy reflecting the topological ordering of the input vector space. We assume a latent similarity in word co-occurrences and use TreeGCS to estimate word similarity from contextual statistics without resorting to a human generated thesaurus. We categorise all keywords as discussed previously and perform no dimensionality reduction thus decreasing information loss. The process is fully automated, requires no human intervention or data processing as the context vectors are generated automatically from unstructured text data and the clustering requires minimal knowledge of the data distribution due to the self-organising network. Each node in the hierarchy represents a small group of synonyms at the lowest level and progressively larger groups of related words up through the tree. The distance between the nodes in the tree is directly proportional to the similarity of the word sets they represent.

3.1. Pre-processing

All upper-case letters are converted to lower-case to ensure matching. A list of all words and punctuation marks in the text corpus is generated and a unique

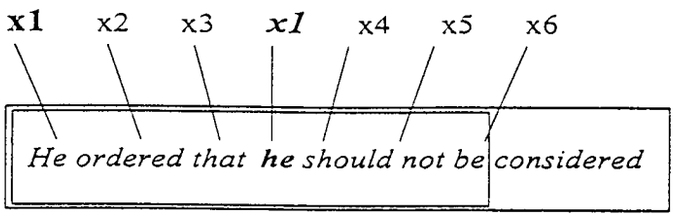


Fig. 1. The moving word window: The initial capital letter will be converted to lower case to ensure the 'he's match. Both instances of 'he' are represented by the same vector. The vectors associated with each word are concatenated to form the context vector for the target word 'he'.

random, real-valued, unit-length m-dimensional vector $\vec{x}$ is assigned to each word as below:

$$Word \rightarrow \vec{x} \in \Re^m. \tag{1}$$

Stop-words are removed to create a second list of keywords. A moving window of size n is passed across the text corpus, one word at a time (see Fig. 1). Ritter and Kohonen use a context window of size three, we use size seven and illustrate the qualitative improvement this generates in Sections 4 and 6. If the word in the window centre is a keyword ($\vec{x}^{\,middle} \in \{keyword\}$), then the unique, random, real-valued, unit length m-dimensional vectors representing each word in the window of size n ($\vec{x}^{\,1} \ldots \vec{x}^{\,n}$) are concatenated and added to the $m * n$ dimensional context vector $\vec{y}_{keyword}$ representing the keyword

$$\vec{y}^{\,keyword} \in \Re^{m*n} = \vec{y}^{\,keyword} + \vec{x}^{\,1} \ldots \vec{x}^{\,n} \wedge (\vec{x}^{\,middle} \in \{keyword\}). \tag{2}$$

When the entire corpus has been processed, all context vectors generated for each keyword are averaged (total for each dimension/frequency of keyword), see below:

$$\vec{Avg}^{\,y} = \forall i \frac{\vec{y}_i}{frequency}, \tag{3}$$

$$\vec{Avg}^{\,y} = symFact * \vec{y}_i \text{ for keyword attributes.} \tag{4}$$

The keyword attributes in the average vector are finally multiplied by the symbol factor (*symFact* in Eq. (4)). The keyword is multiplied by a symbol factor of value 0.2 in Ritter and Kohonen's method for average context vector generation and also in the WEBSOM average context vector generation technique. The symbol factor diminishes the relative influence of the keyword (the central word in the context window) in relation to the surrounding words in the context window for the average context vectors. This prevents the actual keyword over-influencing the topological mapping formation and places the emphasis for topology and semantic similarity inferral on the context vector attributes, the surrounding words. We empirically determined the optimum factor value for our approach with a context window size of seven and found for seven-dimensional vectors that a symbol factor of

0.4 produced the optimal cluster quality (as judged by the authors). For 90-dimensional vectors, the symbol factor has far less influence over the Euclidean distances between the context averages and thus the clusters generated, so we elected to use a symbol factor of 0.4, as this was more effective for the seven-dimensional vectors. This prevents the keyword over-influencing the context average but still provides sufficient influence for a context window of size seven.

As with Deerwester [4], we handle synonymy but only partially accommodate polysemy. Polysemic words are again represented by a weighted average of their contexts but we only generate one context for polysemic words (the context is the mean context of all word senses biased by the frequency of occurrence of each sense). For example, *plant* may be *a living organism* or *heavy machinery*. Only one context average would be produced for *plant*.

3.2. GCS algorithm

Our TreeGCS method is based on the growing cell structure (GCS) method that is described next and is adapted from [5]. GCS networks form discrete clusters unlike SOMs where the SOM cells remain connected in a lattice structure. The dimensions of the SOM lattice have to be pre-specified (such as the 9×9 grid used in our evaluation later in this paper). Contrastingly, only the maximum number of cells needs to be pre-specified in GCS and the network grows dynamically by adding new cells and deleting superfluous cells until the maximum number of cells is reached. The initial topology of GCS is a two-dimensional structure (triangle) of cells (neurons) linked by vertices. We use a two-dimensional cell network to allow a hierarchy to be superimposed and to allow visualisation if necessary. Each cell has a *neighbourhood* defined as those cells directly linked by a vertex to the cell. The input vector distribution is mapped onto the cell structure by mapping each input vector to the best matching cell. Each cell has a $contextWindow * wordVectorDimensionality$—dimensional vector attached denoting the cell's position in the input vector space; topologically close cells have similar attached vectors. On each iteration, the attached vectors are adapted towards the input vector. The adaptation strength is constant over time and only the *best matching unit* (*bmu*) and its direct topological neighbours are adapted unlike SOMs where the adaptation occurs in a progressively reducing radius of neurons around the *bmu*. Cells are inserted where the cell structure under-represents the input vector distribution and superfluous cells that are furthest from their neighbours are deleted. Each cell has a '*winner counter*' variable denoting the number of times that cell has been the *bmu*. The winner counter of each cell is reduced by a pre-determined factor on every iteration. The aim of the GCS method is to evenly distribute the winner counter values so that the probability of any cell being a *bmu* for a random input is equal, i.e., the cells accurately represent the input space.

The GCS learning algorithm is described below, the network is initialised in point 1 and points 2–7 represent *one iteration*. An *epoch* constitutes one iteration (points 2–7) for each input vector in the dataset, i.e. one pass through the entire

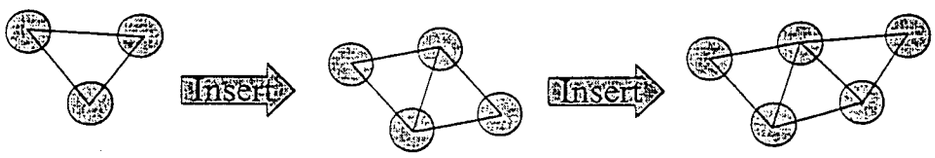


Fig. 2. Cell insertion: A new cell and associated connections are inserted at each step.

dataset.

(1) A random triangular structure of connected cells with attached vectors ($w_{c_i} \in \Re^n$) and E representing winner counter (the number of times the cell has been the winner) is initiated.

(2) The next random input vector ξ is selected from the input vector density distribution. The input vector space is represented as real-valued vectors of identical length.

(3) The best matching unit (*bmu*) is determined for ξ and the *bmu*'s winning counter is incremented.

$$bmu = \|\xi - w_c\|_{min_{c \in network}} \quad \text{where } \| \| \text{ is the Euclidean distance,}$$

$$\Delta E_{bmu} = 1.$$

(4) The best matching unit and its *neighbours* are adapted towards ξ by adaptation increments set by the user.

$$\Delta w_{bmu} = \varepsilon_{bmu}(\xi - w_{bmu}),$$

$$\Delta w_n = \varepsilon_i(\xi - w_n) \quad (\forall n \in neighbourhood).$$

(5) If the number of input signals exceeds a threshold set by the user a new cell (w_{new}) is inserted between the cell with the highest winning counter (w_{bmu}) and its farthest *neighbour* (w_f)—see Fig. 2,

The weight of the new unit is set according to:

$$w_{new} = (w_{bmu} + w_f)/2.$$

Connections are inserted to maintain the triangular network configuration. The winner counter of all *neighbours* of w_{new} is redistributed to donate fractions of the neighbouring cells' winning counters to the new cell and spread the winning counter more evenly,

$$\Delta E_n = -\frac{1}{|n|}E_n \quad (\forall n \in neighbourhood\ of\ w_{new}).$$

The winner counter for the new cell is set to the total decremented from the winning counters of the neighbouring cells.

$$E_{new} = \sum \left(\frac{1}{|n|}E_n \quad (\forall n \in neighbourhood\ of\ w_{new})\right).$$

Fig. 3. Cell deletion: Cell A is deleted. Cells B and C are within the neighbourhood of A and would be left dangling by the removal of the five connections surrounding A, so B and C are also deleted.

(6) After a user-specified number of iterations, the cell with the greatest mean Euclidean distance between itself and its *neighbours* is deleted and any cells within the neighbourhood that would be left 'dangling' are also deleted (see Fig. 3). Any trailing edges are deleted to maintain the triangular configuration.

$$Del = \max_{c \in network} \left(\frac{\sum \|w_c - w_n\|}{card(n)} \; \forall n \in neighbourhood \right).$$

(7) The winning counter variable of all cells is decreased by a user-specified factor to implement temporal decay.

$$\Delta E_c = -\beta E_c \quad \forall c \in network.$$

The user-specified parameters are: the dimensionality of GCS which is fixed at 2 here, the maximum number of neighbour connections per cell, the maximum cells in the structure, ε_{bmu} the adaptation step for the winning cell, ε_i the adaptation step of the neighbourhood, β the temporal decay factor, the number of iterations for insertion and the number of iterations for deletion.

The algorithm iterates until a specified performance criterion is met, such as the network size. If the maximum number of epochs and the maximum number of cells are specified as the termination criteria then new cells are inserted until the maximum number of cells is reached. Once the maximum has been reached, adaptation continues each iteration and cells may be deleted. The cell deletion reduces the number of cells to below the maximum allowing one or more new cells to be inserted until the maximum number of cells is reached again. Deletion removes superfluous cells while creating space for new additions in under-represented regions of the cell structure so the input distribution mapping is improved while the maximum number of cells is maintained.

3.3. TreeGCS algorithm

The TreeGCS is superimposed onto the standard GCS algorithm exposited above. A tree root node points to the initial cell structure and incorporates a list of all cells from the GCS. As the GCS splits or clusters are deleted, the tree divides

and removes leaf nodes to parsimoniously summarise the disjoint network beneath and the GCS cell lists are updated with each leaf node holding a list of all GCS cells in its associated cluster. Only leaf nodes maintain a cluster list. A parent's cluster list is implicitly a union of the children's cluster lists and is not stored for efficiency—minimising memory usage. No constraints are imposed on the tree, hence it is dynamic and requires no prior data knowledge—the tree progressively adapts to the underlying cell structure. The hierarchy generation is run once after each GCS epoch. The running time per hierarchy generation iteration is $O(cells)$ as we essentially breadth first search through the entire cell structure.

A conceptual hierarchy of word synonym clusters is generated. The distance in the hierarchy between two concepts is inversely proportional to the similarity. Concepts are progressively more general and the cluster sets become larger towards the root of the hierarchy.

The underlying GCS's algorithm is susceptible to the ordering of the input vector space, if we alter the order of the input vectors in the dataset, a different cluster topology is generated for each unique input vector order [7]. Thus, in TreeGCS we only commence cell deletion once 90% of the total cells required in the cell structure have been added [7]. This delayed deletion prevents premature cluster committal and ensures the GCS network has evolved sufficiently before cluster splitting commences. In addition, we also iterate between different orders of the input vector space to ameliorate the order susceptibility (the x-dimensional vectors that represent the context averages are rearranged to generate different data orders). Iterating between different orders cancels out the variance in the hierarchical structures generated by the different orders, vastly improving the algorithm's qualitative performance. The algorithm for the tree superimposition is detailed below in pseudocode.

For each epoch,
  Execute the GCS epoch, forming an unconnected graph representing the disjoint clusters.
  Breadth first search from the final winning cell for the epoch to determine which cells are present in the cluster.
  While some cells remain unprocessed,
    Breadth first search from the next unprocessed cell to determine which cells are present in the cluster.
  If the number of clusters has increased from the previous epoch, **then** any tree nodes that point to multiple clusters are identified and child nodes are added for each new cluster formed (see Fig. 4). The cluster list of the parent is deleted and cluster lists are updated for the child nodes. If a cluster is formed from new cells (cells inserted during the current epoch) then a new tree node is added as a child of the root and the new cluster cells added to the new node's list.
  Elsif the number of clusters has decreased, a cluster has been deleted and the associated tree node is deleted. The tree is tidied to remove any redundancy (see Fig. 5).

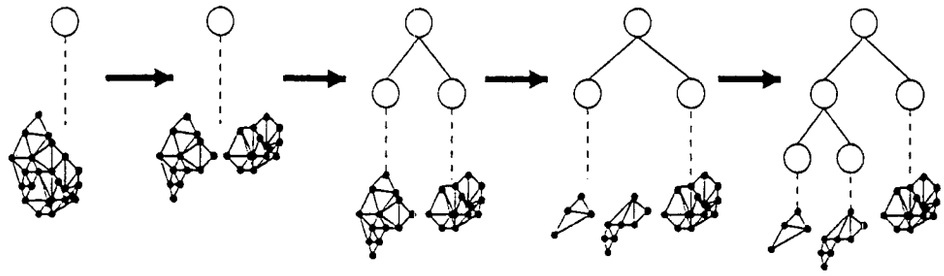


Fig. 4. Cluster subdivision: One cluster splits to form two clusters and the hierarchy is adjusted. The leftmost cluster then splits again.

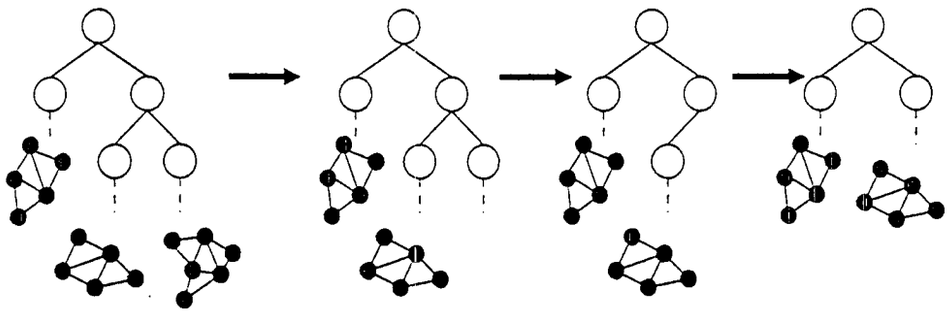


Fig. 5. Cluster deletion: The rightmost cell cluster is deleted during an epoch (step 2)—this leaves a dangling pointer. The node with the dangling pointer is removed (step 3), leaving redundancy in the hierarchy. The redundancy is removed in the final step.

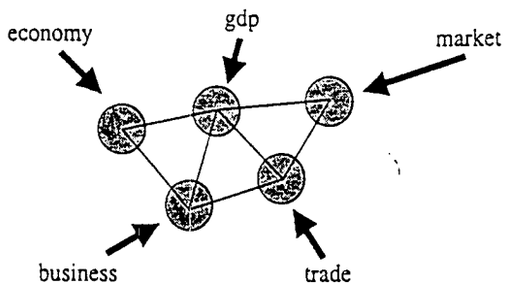


Fig. 6. The cells in the GCS layer are labelled with the words they represent.

For each unprocessed cluster, the tree node that points to that cluster is determined, the cluster list emptied and the new cells are added.

The GCS cells are labelled, see Fig. 6. Each input vector is input to the GCS and the cell identifier of the *bmu* is returned. The cell can then be labelled with

the appropriate word. Words that occur in similar contexts map to topologically similar GCS cells, thus reflecting syntactic and semantic similarity through purely stochastic background knowledge. Tree nodes are merely pointers to GCS cells. All nodes except leaf nodes have only an identifier and pointers to their children. The leaf nodes have an identifier but also point to the GCS cell clusters and implicitly the GCS cell labels (they maintain a list of the identifiers of the GCS cells in their respective clusters). When the GCS *bmu* is identified, the associated tree node can also be identified and the tree can be traversed to find all word distances from the distances between the clusters (leaf nodes) in the tree.

4. Evaluation

We initially demonstrate the qualitative effectiveness of our average vector generation method against the R & K and WEBSOM approaches. We also demonstrate the qualitative effectiveness of our TreeGCS algorithm against the SOM algorithm. Human clustering is innately subjective. In an experiment by Macskassy et al. [16], no two human subjects produced 'similar' clusters when clustering the information contained in a set of Web pages. This creates difficulties for cluster set evaluation and determining whether computerised methods are qualitatively effective. We evaluate the quality of the three methodologies for average context vector generation by comparing the TreeGCS clusters produced from the vectors for each methodology against a dendrogram cluster set produced from the same vectors to provide a Euclidean distance-based evaluation. We then compare the TreeGCS hierarchies against the cluster sets of a human generated thesaurus. We compare TreeGCS and SOM clustering by evaluating the topologies produced for each set of average context vectors against the clusters produced by a dendrogram trained on the same vectors. Fritzke has previously demonstrated GCS's superior performance with respect to correctly classified test patterns over six common neural network approaches and the nearest neighbour statistical classifier for mapping the vowel recognition dataset [5]. In this paper, we use a small dataset comprising 51 words to enable visualisation of the cluster structures and cluster contents and to permit a qualitative comparison of the cluster structures and cluster contents. A larger dataset would preclude visualisation of the cluster structures as they would be too complex to draw and a qualitative comparison of the cluster structures generated would thus be extremely difficult for a larger dataset.

4.1. Three methods for context vector generation

We emulate the *Ritter & Kohonen methodology* as faithfully as possible. We remove common words, punctuation and numbers from the text corpus. We select the vectors from a distribution of random numbered, seven-dimensional vectors. We use a context window of size three. We multiply the keyword vector by a symbol factor of 0.2. The following cluster topologies were generated from the text corpus using words that occurred 10 times or more. We chose to only cluster

word frequency 10 words to ensure the context vectors were truly averaged and not biased by limited exposure and also to eliminate infrequent terms as R & K.

WEBSOM, the new development of the R & K approach, uses 90-dimensional real-valued random vectors for the words. Kaski [10] showed that the orthogonality is proportional to vector dimensionality and we have found that for seven-dimensional vectors, the actual vector assigned to each word in the corpus affects the context averages and thus the similarity and clustering produced. The seven-dimensional approach is also more susceptible to the symbol factor as the multiplier has more effect on the Euclidean distance than for 90-dimensional where the effect is less. WEBSOM extends R & K and uses 90-dimensional word vectors, context window of size three and symbol factor 0.2.

Our methodology described in Section 3 varies slightly from the previous two. We only remove numbers from the corpus, the previous two methods also remove common words and punctuation. We do not generate context averages for common words and punctuation but use them in the context window of other words, hence we have a larger context window of size seven. Our method uses 90-dimensional vectors and symbol factor 0.4. Again, only words occurring 10 or more times are shown in the clusters to ensure the contexts were averaged and not biased by infrequency due to the small size of our test corpus. Although normally we would include these words, we wanted a valid comparison to the previous methods.

We use the Ritter & Kohonen method, WEBSOM method and our method for average context vector generation to produce three sets of vectors. We train each of the three average context vector sets in turn into a standardised benchmark Euclidean distance-based clustering algorithm, the *dendrogram*, to derive three TreeGCS and three dendrogram clusterings, one for each context vector generation method. We compare the TreeGCS and dendrogram cluster topologies for each vector methodology. We produced one 25×25 matrix for each average vector methodology indexed by the 25 most similar words from the dendrogram. The 25 words are shown in bold text in Figs. 7, 9 and 11. For each TreeGCS cluster (Figs. 7, 9 and 11) we placed a 1 at position ij in the respective matrix if $word_i$ and $word_j$ co-occurred in a TreeGCS cluster, otherwise, we entered a 0 in the matrix if they were in different clusters. We filled the half matrix where $i<j$ so there was only one entry per ij pair to prevent redundancy as co-occurrence is symmetric, if ij co-occur then ji must co-occur. After completing each matrix, we counted the number of 1s entered. This represents the number of words clustered together in the dendrogram top 25 cluster that are also clustered together in the TreeGCS cluster structure generated from each vector methodology. The highest score indicates that the vectors generated from that method enable the TreeGCS to most closely emulate the Euclidean distance-based cluster of the dendrogram. The three TreeGCS clusters in Figs. 7, 9 and 11 vary in depth so for a consistent comparison we reduced the depth of the WEBSOM and our TreeGCS trees to level 4 shown in Figs. 9 and 11 which is equivalent to the shallowest hierarchy by merging all clusters below level 4 to form a single leaf cluster at level 4.

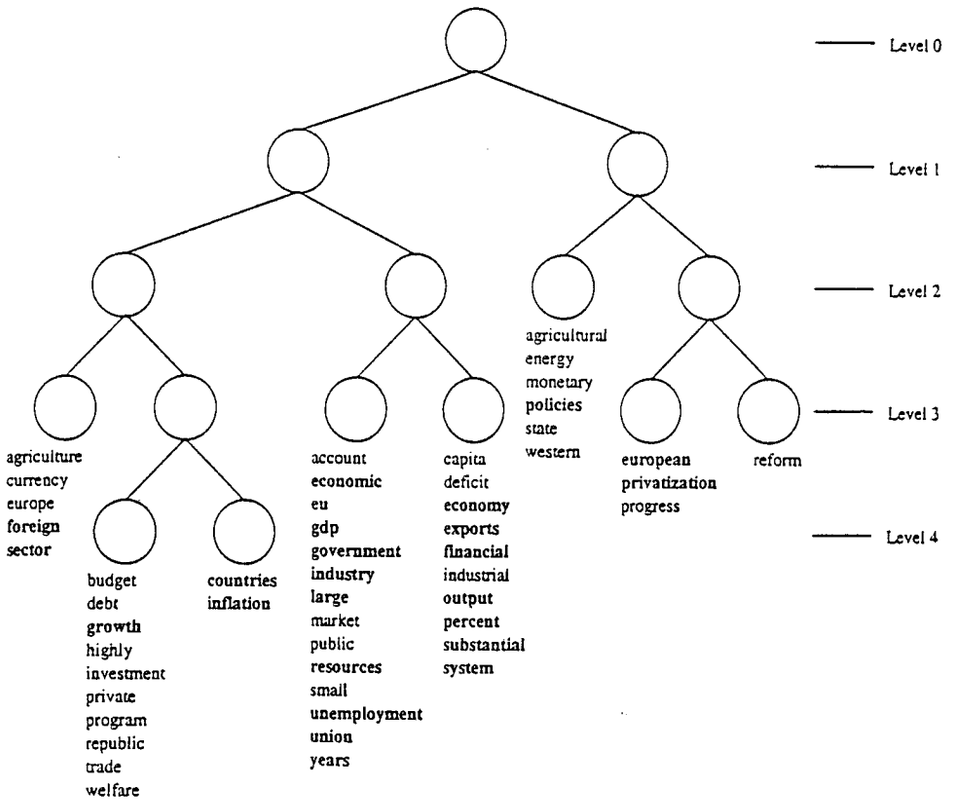


Fig. 7. Ritter & Kohonen methodology—TreeGCS cluster. The figures in bold indicate the top 25 words selected by the dendrogram in each cluster.

We further compare each of the three TreeGCS clusters to a human word hierarchy derived from the MS Bookshelf[2] thesaurus. We produced a 51×51 matrix for the thesaurus clusters and the three TreeGCS hierarchies, indexed by all 51 words in the evaluation. Again, we entered a 1 at position ij in the respective matrix if $word_i$ and $word_j$ co-occurred in a cluster otherwise we entered a 0 in the matrix. Again, we filled the half matrix where $i<j$. This produces four matrices. We overlaid each of the three TreeGCS matrices in turn over the human matrix and counted the number of 1s in the same position in both matrices. This represents the number of words clustered together in both the human and TreeGCS cluster structures, the higher the score, the more similar the TreeGCS clusters are to the human clusters. Again, we repeated the evaluation with the WEBSOM and our TreeGCS hierarchies limited to level 4 for consistency with the R & K hierarchy.

[2] We were unable to use the WORDNET hierarchy as it does not contain all the words from the text corpus. This precludes the use of synSet distances from the WORDNET hierarchy [14] (described previously) as an evaluation tool. Bookshelf allows us to generate clusters distances but no word similarity distances.

4.2. TreeGCS versus SOM clustering comparison

We then train the three sets of average context vectors generated by the three methods into an SOM and TreeGCS for comparison of the accuracy of the two clustering algorithms. We compare TreeGCS versus SOM purely on vector distances by analysing the distribution of the dendrogram words (the 25 closest words with respect to Euclidean distance) through the TreeGCS and SOM clusters. The dendrogram cluster sets act as benchmarks, to ensure that the mapping of input vectors to cells for both TreeGCS and SOMs are preserving the vector distances. We validate that the TreeGCS clusters more accurately emulate the dendrogram cluster sets.

4.3. Text corpus, dendrogram and thesaurus

The text corpus for the evaluation was taken from the economic data in the World Factbook [21] for each of the countries in Europe. This corpus is written in correct English, the vocabulary is reasonably small allowing a compact thesaurus to be generated with many words that have similar meanings allowing the cluster quality to be readily evaluated. We cluster the context averages of the words that occur 10 times or more in the text corpus for all evaluations. This emulates the R & K and WEBSOM methodologies that remove infrequent terms and it maintains a consistency of words to be clustered to ensure a valid and consistent comparison.

The dendrogram hierarchically illustrates similarities and is ideal for structure comparison. The dendrogram uses the centroid-clustering algorithm where the algorithm iteratively merges clusters. Initially, there is one data point per cluster. Each cluster is represented by the average of all its data points, the mean vector; the inter-cluster distance is defined as the distance between pairs of mean vectors. The algorithm iteratively determines the smallest distance between any two clusters, (using the Euclidean distance metric) and the two clusters are merged producing a branch in the cluster tree. The merging is repeated until only one cluster is left. However, dendrograms have problems with identical similarities as only two clusters may be merged at each iteration, so if there are two pairs of clusters with equal distances, one pair has to be merged on one iteration and the other pair on the next iteration, the order being arbitrary. In dendrograms, visualisation is difficult for a large dataset as there is one leaf node for each data point so it is very difficult to view more than 500 data points. Therefore, we feel a dendrogram would be unsuitable as the underlying mechanism for a lexical clustering method but is relevant for structure and cluster comparisons on a small dataset. Both TreeGCS and SOMs use Euclidean distance when mapping the inputs on to the output topology so we feel the dendrogram is consistent with these approaches. Each input vector is represented by a leaf node in the dendrogram. In the SOM and TreeGCS, many vectors can map to leaf nodes so we can use the comparison with the dendrogram to ensure the vector mappings are not distorted and Euclidean distance-based vector similarities preserved when multiple vectors map to leaf nodes.

Table 1
Settings for the context vector generation in each of the three methods evaluated

Method	Vector dimensionality	Context window	Symbol factor
R & K	7	3	0.2
WEBSOM	90	3	0.2
Ours	90	7	0.4

We produced synonym sets from the MS Bookshelf thesaurus to allow comparisons of the clusters generated in our evaluations with a human generated clustering. The synonym sets are arranged in similarity order, the closer together the more similar and the greater the distance the more dissimilar are the words. The synonym sets are:

- {economy, system, market, budget, policies, program, government, account},
- {investment, resources, welfare, privatization, reform},
- {output, energy, exports, gdp, trade},
- {industry, agriculture},
- {growth, progress, inflation},
- {debt, deficit},
- {economic, financial, industrial, monetary},
- {agricultural},
- {currency},
- {capita, percent, sector},
- {substantial, large, highly},
- {small},
- {foreign, private, public},
- {countries, republic, state},
- {european, eu, europe, union, western},
- {unemployment},
- {years}.

4.4. Settings

All settings are summarised in Tables 1–3. Table 1 compares the settings for the generation of the averaged context vectors from the word contexts in the text corpus for each of the three methods evaluated. Table 2 compares the settings for the SOM for each method of word context vector generation. We use the SOM-PAK [13] SOM implementation (as used in WEBSOM [8]). We use the parameter settings that produced the minimal quantisation error for a 9×9 map of rectangular topology, using the neighbourhood kernel 'bubble', (where the neighbourhood function refers to the set of array points around the node). WEBSOM required a different setting for α (the cell vector adaptation parameter) compared to the other two methods to minimise the quantisation error of the topological mapping from the input space to the 9×9 map. Table 3 compares the settings for the TreeGCS for each method

Table 2
Parameter settings for the SOM algorithm to generate the map for each of the three context vector generation methods evaluated[a]

Method	α for x epochs	Radius for x epochs	α for next y epochs	Radius for next y epochs
R & K	0.9 for 3000	10 for 3000	0.5 for 27,000	3 for 27,000
WEBSOM	0.75 for 3000	10 for 3000	0.5 for 27,000	3 for 27,000
Ours	0.9 for 3000	10 for 3000	0.5 for 27,000	3 for 27,000

[a]α is the initial learning rate parameter which reduces to 0 during training and the *radius* is the neighbourhood of cells that are adapted in the SOM adaptation phase. The radius iteratively reduces to 0 during training.

Table 3
Parameter settings for the TreeGCS algorithm to generate the cluster hierarchy for each of the three context vector generation methods evaluated

Method	ε_{bmu}	ε_i	β	Cells	Max. Conns[a]	Insertion	Deletion	Epochs
R & K	0.1	0.01	0.001	81	25	10	810	30,000
WEBSOM	0.02	0.002	0.0002	81	25	10	810	30,000
Ours	0.02	0.002	0.0002	81	25	10	810	30,000

[a]Conns is an abbreviation for connections.

of word context vector generation. We set the parameters to produce the 'best' quality clusters as judged by the authors, see [7] for a discussion of selecting parameter combinations. The seven-dimensional vector evaluation required different parameters from the 90-dimensional trial.

5. Results

We detail the cluster topologies produced by TreeGCS, SOM and the dendrogram for each context vector methodology.

5.1. Ritter & Kohonen method

- From the dendrogram clustering using the R & K average context vectors, the 25 most similar words are:
  {system union output industry substantial policies exports european privatization countries sector inflation percent economic foreign financial resources government growth large economy unemployment gdp years eu}. These words index the 25×25 matrix used to evaluate the average context vector methodologies.
- For the TreeGCS hierarchy generated using the R & K average context vectors, see Fig. 7. The words in bold are the 25 most similar words identified by the dendrogram generated using the R & K average context vectors and are used to form the 25×25 matrix with ij set to 1, where $word_i$ and $word_j$ are in the same TreeGCS cluster.

monetary		program		inflation output		republic trade		
	reform		economy growth system		percent substantial		budget welfare	
		deficit				highly sector		agriculture currency market
	industrial		eu member		europe investment			
european privatization progress		financial		account		debt		
economic			government public		exports industry union			
		large resources		gdp years		agricultural policies		private
			foreign small		low unemployment		state	
capita						countries energy		western

Fig. 8. Ritter & Kohonen methodology—SOM mapping. The words in bold indicate the top 25 words selected by the dendrogram.

- For the SOM cluster topology (see Fig. 8), again the 25 most similar words from the dendrogram are highlighted in bold.

5.2. *WEBSOM*

- From the dendrogram generated using the WEBSOM average context vectors, the 25 most similar words are:
  {countries european budget exports industry sector agricultural industrial large system trade output gdp financial eu economic economy government growth inflation percent privatization unemployment years foreign}. These words form the indices for the 25×25 matrix.

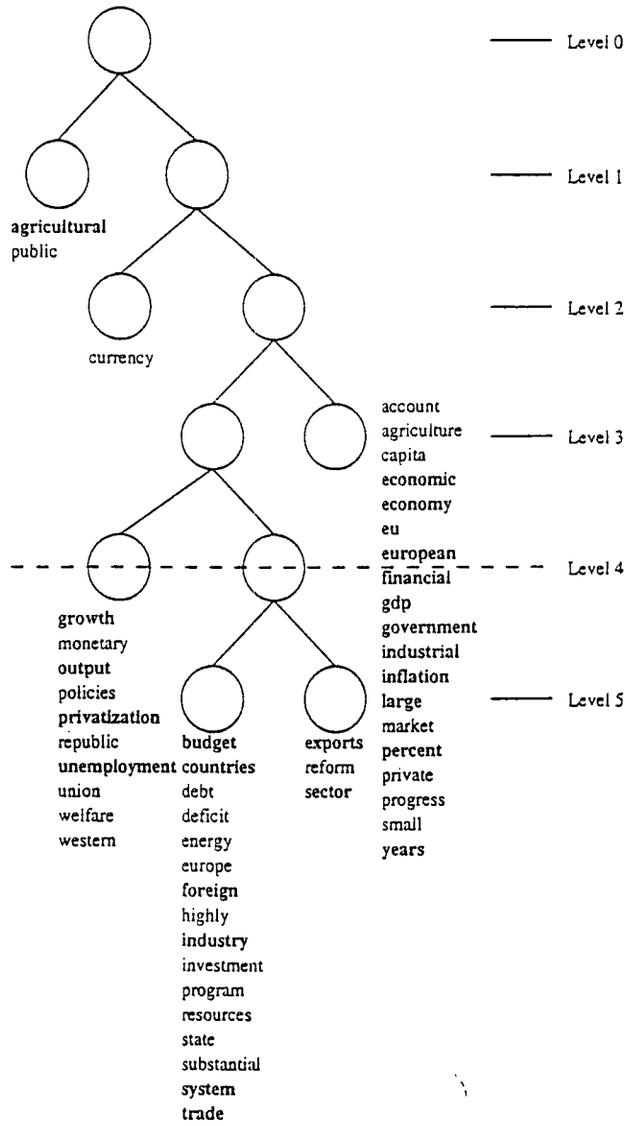


Fig. 9. WEBSOM methodology—TreeGCS cluster. The figures in bold indicate the top 25 words selected by the dendrogram in each cluster.

- See Fig. 9 for the TreeGCS hierarchy generated using the WEBSOM average context vectors. The words in bold are the 25 most similar words identified by the dendrogram generated using the WEBSOM average context vectors.
- For the SOM, the topology is illustrated in (see Fig. 10), again the 25 most similar words from the dendrogram are highlighted in bold and form the co-occurrence entries in the 25×25 matrix.

capita		deficit				progress		agriculture market welfare
	eu						industrial	
currency debt investment trade		exports republic sector		european large		growth percent privatization		agricultural public
	economy		economic		government		financial	
foreign resources				gdp output unemployment		energy industry		program
	budget				inflation			
account union		countries europe		highly system		private small substantial		years
			state					
		policies		western		low		monetary reform

Fig. 10. WEBSOM methodology—SOM mapping. The words in bold indicate the top 25 words selected by the dendrogram.

5.3. Our methodology

- For the dendrogram generated using average context vectors produced by our method, the cluster of the 25 most similar terms is:
  {small reform percent exports gdp output system agricultural market budget industrial financial foreign large industry privatization inflation growth economic eu economy government trade unemployment energy}. These words index the matrix.
- For the TreeGCS hierarchy generated from average context vectors produced by our method, see Fig. 11. The words in bold are the 25 most similar words

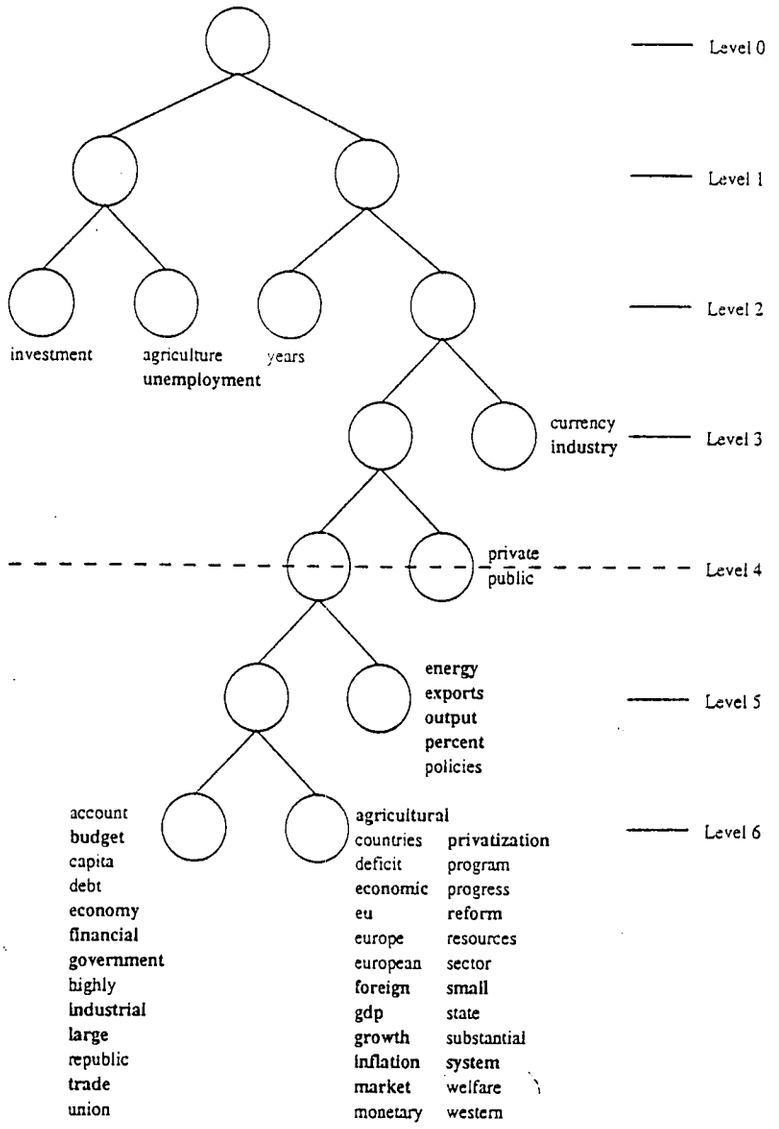


Fig. 11. Our methodology—TreeGCS cluster. The figures in bold indicate the top 25 words selected by the dendrogram in each cluster.

identified by the dendrogram generated from the average context vectors produced by our method. The co-occurrence statistics form the matrix entries.

- For the SOM, the cluster topology is shown in Fig. 12. The 25 most similar words from the dendrogram are shown in bold. We have also included the Sammon mapping (see [13]) for the SOM (see Fig. 13). The Sammon mapping maps the n-dimensional input vectors onto two-dimensional points on a plane.

debt deficit republic sector		public		large substantial		eu	european	
					currency			government
percent years		countries europe		resources small		market		budget economy
			growth		energy		industrial	
policies progress		privatization reform		exports		gdp output		agricultural financial
			inflation		trade		state	
industry program		system		agriculture unemployment		economic private		foreign low
					highly		welfare	
monetary		investment		account union		capita		western

Fig. 12. Our methodology—SOM mapping. The words in bold indicate the top 25 words selected by the dendrogram.

6. Analysis

6.1. Three methods for context vector generation

From Table 4, the TreeGCS cluster produced from our method for average context vector generation is most similar to both the dendrogram and the human cluster sets. When we reduce the TreeGCS structure to level 4 for equality with the shallowest TreeGCS structure, our vector generation method is even more similar to both the dendrogram and human clusters.

The TreeGCS structures generated from 90-dimensional vectors emulate human clusterings more closely than dendrograms from the seven-dimensional vectors. The

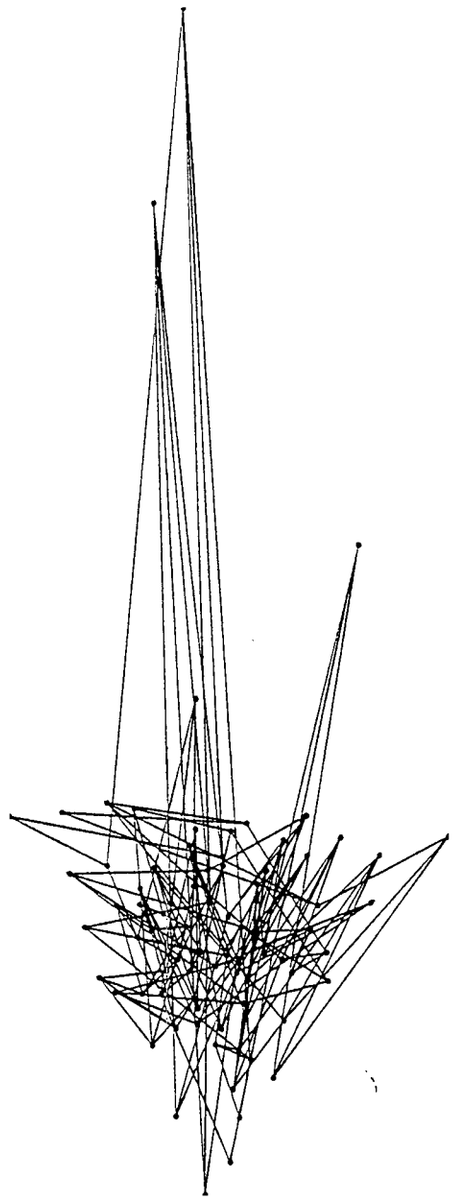


Fig. 13. The Sammon map generated for 90-dimensional vectors with context window being equal to 7.

higher dimensionality vectors increase word vector orthogonality; a prerequisite for the 'bag of words' average context vector generation approach. It is imperative that the vectors ascribed to the individual words in the text corpus imply no ordering of the words, so text processing is based purely on the processing of sequences of

Table 4
TreeGCS clusters produced from each of the three vector generation methodologies against the dendrogram and human generated clusters[a]

Method	Dendrogram	Dendrogram Level 4	Human	Human Level 4
R & K	70	70	10	10
WEBSOM	88	100	22	26
Ours	93	253	32	74

[a]We produced $N \times N$ matrices of all words to be clustered: the 25 most similar words from the dendrograms for the dendrogram comparison and the 51 cluster words for the human comparison. If two words ($word_i$ $word_j$) co-occur in a cluster, then we inserted a 1 in the respective matrix, otherwise we inserted a 0. We then counted the number of 1s in the 25×25 matrix for each vector methodology, where a word pair co-occur in both the dendrogram cluster and the TreeGCS cluster for that vector methodology. The counts are listed in column 2. We reduced all trees to level 4 shown in Figs. 7, 9 and 11 for equality and repeated the evaluation with the counts listed in column 3. We overlaid the human 51×51 matrix against each of the three 51×51 TreeGCS matrices and counted the number of overlaid 1s where a word pair co-occur in both the human and TreeGCS clusters, listed in column 4. Again we repeated the evaluation for the level 4 trees and the counts are given in column 5.

words. For seven-dimensional vectors, the Euclidean distances are altered between the context averages when different vectors are initially assigned to different words. This is particularly important for low frequency words where the context average is biased by the vectors assigned. Even for words occurring more than 10 times, the vector assignment influences the similarities. Kaski [10] showed that there is a direct correlation between vector dimensionality and orthogonality—the higher the dimensionality, the greater the orthogonality. We empirically evaluated various dimensionalities for consistency with respect to cluster content when different vectors are initially ascribed to the words in the corpus. We used the dendrogram to pinpoint the most similar 25 words. We found that the cluster sets were identical for 90-dimensional vectors over a set of experiments but varied for all dimensionalities tested below 90. The higher dimensionality spreads the vectors more across the input pace allowing a more accurate differentiation of clusters. We feel similar methods, using self-organising maps or growing cell structures, should use vectors of this dimensionality or greater to ensure orthogonality and spread and to maintain consistency and stability of the lexical clusters regardless of initial word-vector assignments.

With respect to the size of the context window, we feel that our size seven-context window produces superior quality TreeGCS clusters to WEBSOM's context window of size three. The TreeGCS clusters produced from the average context vectors produced by our method emulate both the nearest neighbour (dendrogram) and human generated thesaurus more accurately than the TreeGCS cluster produced from the WEBSOM average context vectors. The vast majority of terms from the dendrogram and MS Bookshelf are in the three clusters (see Fig. 11) for our vector generation method but are spread across four clusters with many of the other words

also within these clusters for the WEBSOM method of vector generation (see Fig. 9).

6.2. TreeGCS versus SOM clustering comparison

For all three evaluations in Sections 5.1–5.3, the top 25 words from the dendrogram are spread across the SOM (as can be seen from the spread of bold text in Figs. 7, 9 and 11) but tend to be in closely related clusters in the TreeGCS hierarchy with just the odd exception (the bold text occurs in clusters that are near neighbours in the hierarchy). For example, for our method (see Fig. 11), the dendrogram words, shown in bold text, are predominantly in three clusters and these clusters are very closely related. Only 'industry' and 'unemployment' are clustered elsewhere. With respect to Euclidean distance, the TreeGCS emulates the nearest neighbour approach of the dendrogram far better than the SOM. The Sammon mapping produced from the SOM using our method to derive the context vectors is extremely distorted (see Fig. 13). SOMs are criticised in the literature [1] for distorting high-dimensional inputs when they map onto the two-dimensional representation.

7. Conclusion and future work

We feel that our method, 90-dimensional vectors, symbol factor of 0.4, context window of seven is superior to the R & K and WEBSOM methods. Our method for context vector generation enables TreeGCS to be more similar to both the dendrogram (Euclidean distance) and the human generated thesaurus than either the R & K or WEBSOM approaches. We note that the corpus effects the similarity of the computer generated structures against a human thesaurus. The human thesaurus encompasses general word meanings while the text corpus may be very specific so the similarity of the computer generated approaches to the human clusters is affected and may appear low. We also demonstrated that the TreeGCS algorithm emulates Euclidean vector-distance based cluster sets more faithfully than the SOM algorithm. Therefore, we feel the optimum approach for synonym clustering of the methods evaluated is to generate the average context vectors using our method and train these in to the TreeGCS cluster algorithm. TreeGCS not only emulates the nearest neighbour and human generated clusters more faithfully, it forms discrete clusters and dynamically forms a lexical hierarchy.

There are two main drawbacks to our current method. The first is the inability to disambiguate words. All senses of a polysemic word are averaged together during the context average formation, distorting the averaged context vectors produced. We intend to improve this by including part-of-speech tagging to differentiate identical words which represent different parts-of-speech, for example *spring: noun, a water source* and *spring: verb, to jump*. However, autonomously differentiating word senses is currently intractable and relies on a knowledge engineer to tag the senses.

The second main drawback lies in the underlying GCS algorithm and is a speed problem. The algorithm is dependent on the winner search—finding the best matching unit. This involves comparing the input vector to the vector attached to each cell, calculating the difference for each vector dimension. This must be repeated for each vector in the input vector space to complete each epoch. This search is therefore, (number of input vectors * vector dimensionality * number of cells) $\approx O(n^3)$ for each GCS epoch. For the small vocabulary evaluated in this paper the speed problem was not apparent. However, for a large corpus, with an extensive vocabulary the speed is slow and we need to speed the algorithm, reduce the running time and hence remove the bottleneck.

Acknowledgements

We would like to thank the anonymous reviewers of our original draft for their incisive remarks and recommendations which allowed us to improve the overall presentation of the paper.

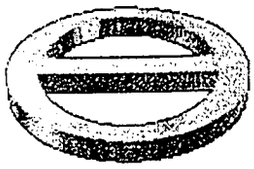
References

[1] C.M. Bishop, Neural networks for pattern recognition, Clarendon Press, Oxford, 1995.
[2] H. Chen, C. Schuffels, R. Orwig, Internet categorization and search: a machine learning approach, J. Visual Comm. Image Representation, Special Issue Digital Libraries 7 (1) (1996) 88–102.
[3] I. Dagan, L. Lillian, F.C.N. Pereira, Similarity-based methods for word sense disambiguation, in: 35th Annual Meeting of the Association for Computational Linguistics, San Francisco, CA, Morgan Kaufmann, Los Altos, CA, 1997.
[4] S. Deerwester, S.T. Dumais, T.K. Landauer, G.W. Furnas, R.A. Harshman, Indexing by latent semantic analysis, J. Soc. Inform. Sci. 1 (6) (1990) 391–407.
[5] B. Fritzke, Growing cell structures—a self-organizing network for unsupervised and supervised learning, Technical Report TR-93-026, International Computer Science Institute, Berkeley, CA, 1993.
[6] V. Hodge, J. Austin, Hierarchical growing cell structures: TreeGCS, in: Proceedings of the Fourth International Conference on Knowledge-Based Intelligent Engineering Systems, August 30–September 1, 2000.
[7] V. Hodge, J. Austin, Hierarchical growing cell structures: TreeGCS, IEEE Transactions on Knowledge and Data Engineering, Special Issue on Connectionist Models for Learning in Structured Domains, 2001.
[8] T. Honkela, S. Kaski, K. Lagus, T. Kohonen, WEBSOM—self-organizing maps of document collections, in: Proceedings of WSOM '97, Workshop on Self-Organizing Maps, Espoo, Finland, June 4–6, Helsinki University of Technology, Neural Networks Research Centre, Espoo, Finland, 1997, pp. 310–315.
[9] T. Honkela, V. Pulkki, T. Kohonen, Contextual relations of words in Grimm tales, analyzed by self-organizing map, in: F. Fogelman-Soulie, P. Gallinari (Eds.), Proceedings of the International Conference on Artificial Neural Networks (ICANN-95), Vol. 2, EC2 et Cie, Paris, 1995, pp. 3–7.
[10] S. Kaski, Dimensionality reduction by random mapping: fast similarity computation for clustering, in: Proceedings of IJCNN '98, International Joint Conference on Neural Network, Vol. 1, IEEE Service Center, Piscataway, NJ, 1998, pp. 413–418.
[11] T. Kohonen, Self-organizing maps, Vol. 2, Springer, Heidelberg, 1997.

[12] T. Kohonen, Self-organization of very large document collections: state of the art, in: L. Niklasson, M. Bodén, T. Ziemke (Eds.), Proceedings of ICANN98, the Eighth International Conference on Artificial Neural Networks, Vol. 1, Springer, London, 1998, pp. 65–74.
[13] T. Kohonen, J. Hynninen, J. Kangas, J. Laaksonen, Som_pak, the self-organizing map program package.
[14] X. Li, S. Szpakowicz, S. Matwin, A WordNet-based algorithm for word sense disambiguation, in: Proceedings of IJCAI-95, Montréal, Canada, 1995.
[15] W. Lowe, Semantic representation and priming in a self-organizing lexicon, in: Proceedings of the 4th Neural Computation Psychology Workshop, Springer, Berlin, 1997, pp. 227–239.
[16] S.A. Macskassy, A. Banerjee, B.D. Davison, H. Hirsh, Human performance on clustering web pages: a preliminary study, in: The Fourth International Conference on Knowledge Discovery and Data Mining, 1998.
[17] F. Pereira, N. Tishby, L. Lee, Distributional clustering of english words, in: Proceedings of ACL-93, Columbus, OH, 1993.
[18] H. Ritter, T. Kohonen, Self-organizing semantic maps, Biol. Cybernet. 61 (1989) 241–254.
[19] H. Schütze, J.O. Pederson, Information retrieval based on word senses, in: Fourth Annual Symposium on Document Analysis and Information Retrieval, Las Vegas, NV, 1995, pp. 161–175.
[20] J. Stetina, S. Kurohashi, M. Nagao, General word sense disambiguation method based on a full sentential context, in: Coling-ACL '98 Workshop "Usage of WordNet in Natural Language Processing Systems", Université de Montréal, Montréal, Canada, August 1998.
[21] World FactBook, http://www.odci.gov/cia/publications/factbook/country-frame.html.
[22] D. Yarowsky, Unsupervised word sense disambiguation rivaling supervised methods, in: Proceedings of the 33rd Annual Meeting of the Association for Computational Linguistics, Cambridge, MA, 1995, pp. 189–196.
[23] G.K. Zipf, Human Behaviour and the Principle of Least Effort: an Introduction to Human Ecology, Addison-Wesley, Cambridge, MA, 1949.

Prof. Jim Austin has the Chair of Neural Computation in the Department of Computer Science, University of York, where he is the leader of the Advanced Computer Architecture Group. He has extensive expertise in neural networks as well as computer architecture and vision. Jim Austin has published extensively in this field, including a book on RAM based neural networks.

Victoria Hodge is a PostGraduate Research Student in the Department of Computer Science, University of York. She is a member of the Advanced Computer Architecture Group investigating the integration of neural networks and information retrieval.



Thesaurus Construction

Welcome to the Introductory Tutorial on Thesaurus Construction

The purpose of this module is to teach you the basics of constructing an information retrieval thesaurus.

Section 1

Contents

Section 1: What Is a Thesaurus?

Section 2: Collecting Terms

Section 3: Modifying and Inventing Terms

Section 4: Preferred Terms and Non-preferred Terms

Section 5: Semantic Relations

Section 6: BT, NT, and RT References

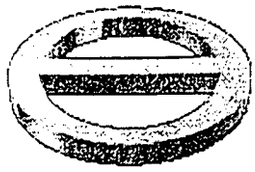
Section 7: Scope Notes

Section 8: Thesaurus Displays

Glossary

Last updated February 19, 2002.
This page maintained by Prof. Tim Craven
E-mail (text/plain only): craven@uwo.ca
Faculty of Information and Media Studies
University of Western Ontario,
London, Ontario
Canada, N6A 5B7

Copyright © 1997 The University of Western Ontario



Section 1
What is a Thesaurus?

A thesaurus is a tool for vocabulary control. By guiding indexers and searchers about which terms to use, it can help to improve the quality of retrieval.

Usually, a thesaurus is designed for indexing and searching in a specific subject area. Examples of subject areas covered by thesauri are education, metallurgy, and art and architecture.

What is in a thesaurus?

A thesaurus gives several types of information to indexers and searchers.

Preferred Terms

Obviously, the thesaurus has to indicate which terms indexers and searchers are allowed to use. These terms are called preferred terms.

This is a major part of vocabulary control - restricting the vocabulary so that it is easier to predict what words might have been used to index a concept.

Non-preferred Terms

In addition to preferred terms, a thesaurus also needs to indicate some terms that indexers and searchers are not to use.

These terms are called non-preferred terms.

It should be possible to look up a non-preferred term and see what preferred term should be used instead. This will save time and make it less likely that the best preferred term will be missed.

A thesaurus also usually allows you to look up a preferred term and see its non-preferred terms. This can give you a better idea of what the term is supposed to mean.

Semantic Relations

As well as linking preferred terms with non-preferred terms, a thesaurus also shows links between different preferred terms.

These links are usually for semantic relations.

Like a link between a preferred term and a non-preferred term, one of these semantic links can help to direct you to the right term and make the meaning of a term clearer.

Guides to Application

A good thesaurus should make it clear what a term is meant to cover. It can accomplish this to some extent by showing non-preferred terms and semantic relations. Other ways of guiding people in using a thesaurus include introductory matter and scope notes.

A scope note often takes the form of a definition of the term.

Ensuring that terms are used consistently with the same meaning is another major aspect of vocabulary control.

Rules for Synthesis

Usually, a thesaurus lists all its preferred terms explicitly.

Such thesauri are enumerative.

Some thesauri indicate some preferred terms indirectly: instead of listing all the preferred terms, they give rules for creating them out of components.

Such thesauri are at least partly synthetic.

Summary

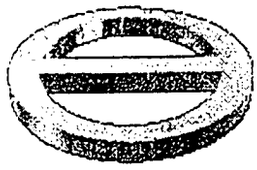
You would expect the following topics to be covered in a thesaurus:

1. preferred terms
2. non-preferred terms
3. semantic relations between terms,
4. how to apply terms
   and sometimes
5. how to produce terms that are not listed explicitly in the thesaurus.

Quiz on thesaurus content (requires JavaScript for Netscape Navigator 3.0)

Introduction Section 2 Table of Contents Glossary

Last updated February 19, 2002.
This page maintained by Prof. Tim Craven
E-mail (text/plain only): craven@uwo.ca
Faculty of Information and Media Studies
University of Western Ontario,
London, Ontario
Canada, N6A 5B7



Section 3
Modifying and Inventing Terms

Standardizing the Form of Words

Terms collected should already be nouns or noun phrases. Here are some further guidelines for the form that terms should take in your final thesaurus.

Guidelines	Examples
Plural for things that can be counted	"TUBES"
Singular for "mass" nouns	"WOOD"
Singular for processes, properties, and conditions	"REFRIGERATION" "WEIGHT", "POVERTY"
Not inverted	"RADAR ANTENNAS" (rather than "ANTENNAS, RADAR")
Excluding prepositions	"CARBOHYDRATE METABOLISM" (rather than "METABOLISM OF CARBOHYDRATES")
Excluding punctuation marks, diacritics, special characters, and abbreviations	"COOPERATIVE PROGRAMS" (rather than "CO-OPERATIVE PROGRAMS" or "COÖPERATIVE PROGRAMS") "MUSICAL NOTES" (rather than "(MUSICAL) NOTES" or "MUS. NOTES"

Quiz on term form (requires JavaScript for Netscape Navigator 3.0)

What to Do with Terms with More Than One Meaning

A homograph is an expression that has the same spelling as another expression, but a different meaning. A thesaurus needs to distinguish between homographs.

A unique term may be created out of a homograph by adding a parenthetical qualifier; for example, "PORT (WINE)".

You may note that including parentheses is contrary to the guideline given above; namely, to avoid punctuation. A unique term may also be created out of a homograph by adding another word without punctuation; for example, "PORT WINE".

Introducing New Terms

In addition to terms extracted from your various sources, you may sometimes choose to introduce new terms of your own.

For example,

1. broad concept terms
2. structural terms
3. terms for nontextual material

Broad Concept Terms

Terms that represent broad concepts may be introduced because they are useful in broad searches.

For example, "TRAFFIC STATIONS", because it can be used to replace a search for "AIRPORTS OR BUS TERMINALS OR TRAIN STATIONS OR HELIPORTS OR...".

Structural Terms

Terms may also be introduced because they help to clarity the structure of semantic relations.

For example, "EMPLOYMENT OF SPECIFIC GROUPS" to clarify the relationship between "EMPLOYMENT" and "YOUTH EMPLOYMENT".

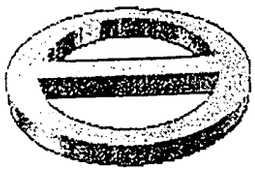
Terms for Nontextual Material

If you are constructing a thesaurus for indexing material which is not in the form of text, you have fewer sources for terms. You may therefore find yourself inventing your own terms more.

For example, this picture shows things that are not referred to in the caption, including a bottle. After examining the picture, you might decide to add "BOTTLES" to your list of terms.

Section 2 Section 4 Table of Contents Glossary

Last updated February 19, 2002.
This page maintained by Prof. Tim Craven



Section 4
Preferred Terms and Non-preferred Terms

Equivalent Terms

After collecting terms for your thesaurus, you need to decide which are equivalent terms. For purposes of indexing and searching, a set of equivalent terms will all be treated as though they meant the same thing and will be represented by a single preferred term.

Spelling and Synonyms

Sometimes, equivalent terms really do mean the same thing. So, it obviously makes sense to use a single preferred term to represent that one meaning.

1. A word may have more than one spelling; for example, "AESTHETICS" and "ESTHETICS".
2. Two different words may have essentially the same meaning; for example, "AUTOMATION" and "MECHANIZATION".

Quasi-synonyms

Sometimes, equivalent terms mean different things in ordinary language. For indexing and retrieval, it is better to group the different meanings together. Such equivalent terms are called quasi-synonyms.

Types of quasi-synonyms

Terms with **overlapping meanings** are sometimes treated as equivalent. For example, "GENIUSES" and "PRODIGIES" might be treated as equivalent, even though the two terms mean different things.

A term whose **scope is included in that of another term** is sometimes treated as equivalent. For example, "STEEL" might be treated as equivalent to "METAL" if it is not important to distinguish items on steel from items on other metals.

Sometimes **opposites** are treated as equivalent, because items on one are likely to be relevant to a query for the other. For example, "TRANSPARENCY" might be treated as equivalent to "OPACITY".

Preferred Terms

Preferred terms serve as focal points where all the information about a concept is collected.

Non-preferred Terms

Non-preferred terms are included in a thesaurus mainly to help users find the appropriate preferred terms. Non-preferred terms may also help to define the scope of preferred terms.

USE/UF

A non-preferred term is normally linked to a corresponding preferred term by a USE reference. The corresponding reference in the opposite direction if UF ("Used For").

For example,

PERIODICALS USE **SERIALS**	**SERIALS** UF **PERIODICALS**

Here the preferred term is "SERIALS" and the corresponding non-preferred term is "PERIODICALS".

Choosing Preferred Terms

The following are some principles for choosing preferred terms, together with examples of applying them.

Guidelines	Examples
Usage	**COOKING** UF **COOKERY** ("Cooking" is the more commonly used word.)
Breadth	**PLASTICS** UF **POLYETHYLENE** ("Plastics" clearly means all plastics, of which polyethylene is only one.)
Disambiguation	**AMERICAN LIBRARY ASSOCIATION** UF **ALA** ("ALA" could stand for something else.)
Collocation	**RAILWAY STATIONS** UF **TRAIN STATIONS** (In an alphabetical sequence, "RAILWAY STATIONS" would appear near to "RAILWAYS" and other terms related to railways.)
Conciseness	**MUCKRAKERS** UF **MUCKRAKING MOVEMENT** (One word rather than two.)
Plural for countable objects	**GEESE** UF **GOOSE** (Geese are countable.)
Internal consistency	If you have decided to prefer the Latin names for plants, do so consistently.
External consistency	You might prefer "PIERS & WHARVES" to "LANDINGS", "BOAT LANDINGS", "DOCKS", "QUAYS", or "WHARVES" partly because that is what the Library of Congress *Thesaurus for Graphic Material* does.

Quiz on preferred terms (requires JavaScript for Netscape Navigator 3.0)

Compound USE References

Instead of a single non-preferred term, one may sometimes instruct indexers and searchers to use more than one preferred term in combination. In such cases, the USE reference points to all the preferred terms, and the UF reference is often marked in some special way.

For example,

SNOWMOBILES USE VEHICLES+SNOW
SNOW UF+ SNOWMOBILES
VEHICLES UF+ SNOWMOBILES

You are especially likely to do this if the non-preferred term consists of more than one word.

For example,

SCHOOL CAFETERIAS USE CAFETERIAS+SCHOOLS
CAFETERIAS UF+ SCHOOL CAFETERIAS
SCHOOLS UF+ SCHOOL CAFETERIAS

On the other hand, you may choose not to make such a term a non-preferred term, even if it consists of more than one word.

Making Multi-word Terms Preferred

When should you allow a multi-word as a preferred term?

A term consisting of more than one word should typically be made a preferred term if

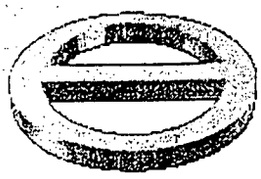
1. combining terms is not possible either at the indexing stage or at the searching stage
2. too many terms would otherwise be required to index an item
3. the resulting number of preferred terms is not too large
4. indexing and searching are generally easier using the compound term
5. the term is likely to be used frequently in indexing or searching

6. the term's components occur frequently in different syntactic relations; for example, "LIBRARY SCHOOLS", "SCHOOL LIBRARIES".
7. the term is needed in the structure of semantic relations; especially, if any narrower concepts are represented by preferred terms.
8. you are in doubt

Section 3 Section 5 Table of Contents Glossary

Last updated February 19, 2002.
This page maintained by Prof. Tim Craven
E-mail (text/plain only): craven@uwo.ca
Faculty of Information and Media Studies
University of Western Ontario,
London, Ontario
Canada, N6A 5B7

Copyright © 1997 The University of Western Ontario



Section 5
Semantic Relations

Why Indicate Semantic Relations?

Indicating semantic relations helps in several aspects of information management:

1. checking **whether a term should be used** in indexing a given item or in formulating a given search specification
2. choosing the correct **level of generality** in indexing and searching
3. searching in response to **broad inclusive queries**
4. **sharing** indexing by facilitating translation from one scheme to another

Semantic Relations between Terms

The main semantic relations indicated between preferred terms in a thesaurus are hierarchical relations and non-hierarchical relations.

BT and NT Links

BT and NT links are used to indicate hierarchical relations. In a hierarchical relation, one term is viewed as being "above" another term because it is broader in scope.

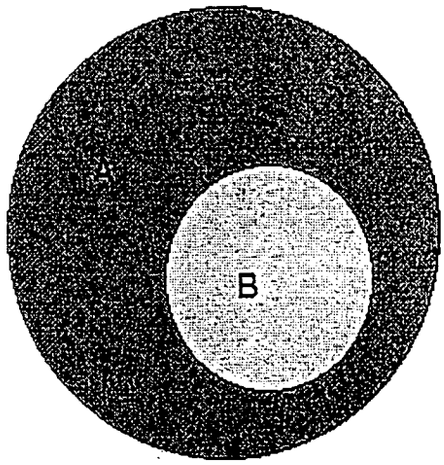
In developing a thesaurus, it is often a good idea to work out the hierarchical relations first.

When Is There a Broader/Narrower Term Relation?

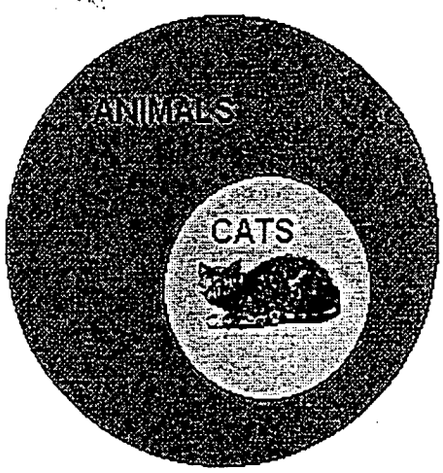
There are various definitions of what constitutes a hierarchical relation. You are advised, however, to restrict yourself to the following cases.

Genus/Species

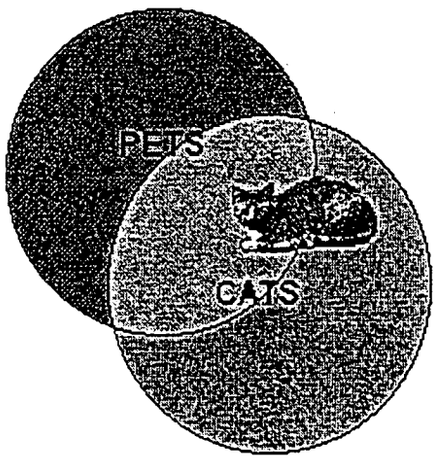
Term A is a broader term to term B (and term B is a narrower term to term A) if all the things included in the class named by term B are included in the class named by term A.



For example, "ANIMALS" is a broader term to "CATS" (and "CATS" is a narrower term to "ANIMALS") because all cats are animals.



On the other hand, "PETS" is not a broader term to "CATS" because not all cats are pets.



Class/Member

The **narrower term** can sometimes name a **class with only one member**.

For example, "UNIVERSITIES" is a broader term to "UNIVERSITY OF WESTERN ONTARIO" because The University of Western Ontario is a university.

Since thesauri usually do not include proper names, you may not encounter cases like this in constructing your own thesaurus.

Hierarchical Whole-Part

Term A is a broader term to **term B** (and term B is a narrower term to term A) if everything included in the class named by term B is a part of something included in the class named by term A.

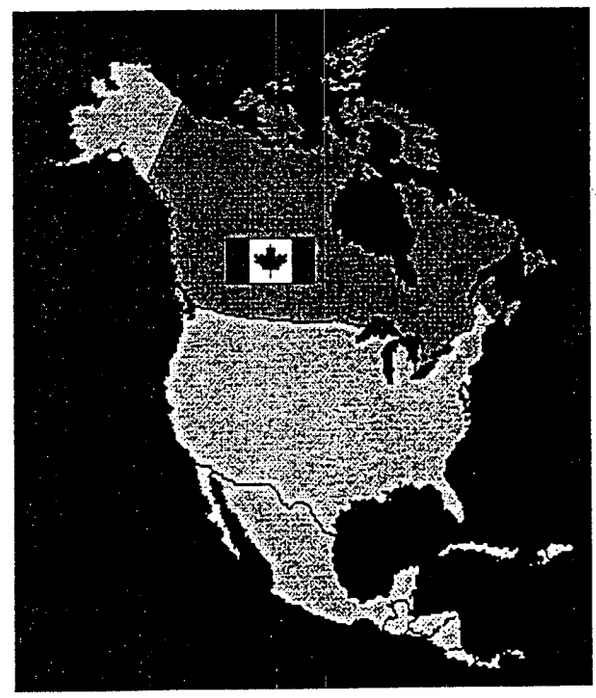
For example, in a medical thesaurus, "HEAD" might be a broader term to "NOSE" because noses are normally parts of heads.

On the other hand, "FORESTS" would not be a broader term to "TREES" because not every tree is part of a forest.

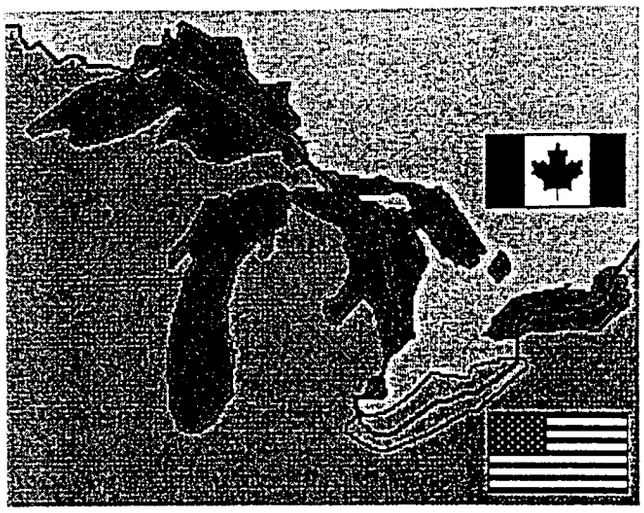
Geographical Whole/Part

In a hierarchical whole/part relation, both the broader term and the narrower term may name a **class with only one member**. This is often true of **geographical names**.

For example, "NORTH AMERICA" is a broader term to "CANADA" because Canada is a part of North America.



On the other hand, "CANADA" is not a broader term to "LAKE ERIE" because only part of Lake Erie is part of Canada.



Since many thesauri do not include geographical names, you may not encounter cases like this in constructing your own thesaurus.

Summary

In recognizing hierarchical relations, you are advised to restrict yourself to the following types:

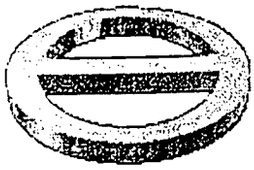
1. **genus/species**
   and its special case
   class/member
2. **hierarchical whole/part**
   and its special case
   geographical whole/part

Quiz on hierarchical relations (requires JavaScript for Netscape Navigator 3.0)

Section 4 Section 6 Table of Contents Glossary

Last updated February 19, 2002.
This page maintained by Prof. Tim Craven
E-mail (text/plain only): craven@uwo.ca
Faculty of Information and Media Studies
University of Western Ontario,
London, Ontario
Canada, N6A 5B7

Copyright © 1997 The University of Western Ontario



Section 6
BT, NT, and RT References

What Is the Relationship Between BT and NT?

Normally, BT and NT are "inverse" links. In other words, if X is a broader term to Y, then Y is a narrower term to X, and vice versa.

For example, if a thesaurus contains the entry

PENS
BT **WRITING MATERIALS**

you would expect it also to have the entry

WRITING MATERIALS
NT **PENS**

How Many BT References Can a Term Have?

A thesaurus is usually "polyhierarchical"; this means that a term can have more than one immediately broader term and more than one BT reference. For example,

SOCIAL PSYCHOLOGY
BT **PSYCHOLOGY**
BT **SOCIOLOGY**

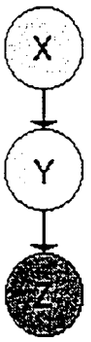
Polyhierarchy avoids futile arguments about the "best" broader term to choose.

Some terms in a thesaurus have no broader terms and so no BT references. Such terms are usually fairly broad in meaning, at least within the subject area covered by the thesaurus. For example, in a sports thesaurus, "SPORTS" might have no broader terms.

When Should BT/NT References Be Omitted?

You should not indicate every hierarchical relation explicitly in your thesaurus. That could make the entries too long and difficult to read. Instead, omit those links that are implied by other links.

Suppose X is a broader term to Y, which in turn is a broader term to Z. Do not make BT/NT references between X and Z.



For example,

PLANT PRODUCTS
NT **FRUIT**

and

FRUIT
NT **FRESH FRUIT**

but not

PLANT PRODUCTS
NT **FRESH FRUIT**

When To Use an RT Reference

An RT reference is used for non-hierarchical semantic relations in a thesaurus. To decide whether there should be an RT reference between two preferred terms X and Y that do not have a hierarchical relation, you can use the following test:

Should an indexer or a searcher considering using X be reminded of the existence of Y?

What Is the Relationship between RT and RT?

Normally, RT is its own "inverse" link type. In other words, if X has an RT reference to Y, then Y should have an RT reference to X. For example, if a thesaurus contains the entry

PENS
RT **CALLIGRAPHY**

you would expect it also to have the entry

CALLIGRAPHY
RT **PENS**

Semantic Categories of RT References

In constructing your thesaurus, you may find it useful to list some categories of semantic relations that you think should be covered by RT references.

Here are some categories sometimes used, with examples.

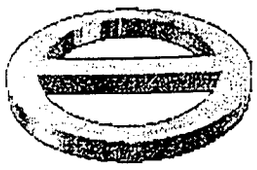
Categories	Examples
Time	**LEISURE READING** RT **LEISURE TIME**
Place	**FOREIGN LANGUAGES** RT **LANGUAGE LABORATORIES**
Product	**STILL CAMERAS** RT **PHOTOGRAPHS**
	SHIPBUILDING RT **SHIPS**
Cause	**VANDALISM** RT **HOSTILITY**
Agent	**COACHING** RT **COACHES**
Device	**PAINTING** RT **PAINT BRUSHES**
Application	**COMPUTERS** RT **WORD PROCESSING**
Part	**VEHICLES** RT **WHEELS**
Complement	**PARENTS** RT **CHILDREN**

Quizzes on thesaurus references (requires JavaScript for Netscape Navigator 3.0)

Section 5 Section 7 Table of Contents Glossary

Last updated February 19, 2002.
This page maintained by Prof. Tim Craven
E-mail (text/plain only): craven@uwo.ca
Faculty of Information and Media Studies
University of Western Ontario,
London, Ontario
Canada, N6A 5B7

Copyright © 1997 The University of Western Ontario



Section 7
Scope Notes

The most common type of guide to applying terms in a thesaurus is the **scope note.** A scope note is normally preceded by the notation SN.

Scope notes take a variety of forms.

Definitions in Scope Notes

A scope note may be a **definition**; for example,

SPACE ERROR
SN TENDENCY TO BE BIASED BY THE SPATIAL POSITION OF STIMULI IN RELATION TO THE OBSERVER

Form of Definitions in Scope Notes

A definition in a scope note should apply to the **noun form,** not to a related verb or adjective.

For example, this scope note for "INDEXING"

SN TO ASSIGN NATURAL LANGUAGE TERMS TO DOCUMENTS

should be changed to

SN ASSIGNING OF NATURAL LANGUAGE TERMS TO DOCUMENTS

Content of Definitions in Scope Notes

A thesaurus term should have a single meaning. Any definitions in the term's scope note should reflect that meaning.

For example, this scope note for "ACCENT"

SN STRESS PLACED ON A SYLLABLE; VARIATION IN PRONUNCIATION DUE TO LINGUISTIC BACKGROUND

is incorrect because it confuses two different meanings of the term.

Including Concepts with Scope Notes

A scope note may indicate a concept that is included in the scope of the term; for example,

MECHANIZED INFORMATION RETRIEVAL
SN INCLUDES PRE-COMPUTER METHODS, SUCH AS PUNCHED CARD SYSTEMS

Excluding Concepts with Scope Notes

A scope note may indicate a concept that is excluded from the scope of the term.

This may be done to show that the term has a narrower meaning than some users of the thesaurus might have in mind; for example,

BEARS
SN DOES NOT INCLUDE PANDAS

It may also be done to draw attention to an excluded meaning of an ambiguous term; for example,

PARTIES
SN POLITICAL PARTIES ONLY. DO NOT USE FOR SOCIAL GATHERINGS

References to Other Terms in Scope Notes

Some scope notes refer to other terms, especially to indicate how to deal with a concept that is excluded; for example,

LICENSING
SN EXCLUDES ASPECTS COVERED BY THE TERMS 'SCHOOL ACCREDITATION' AND 'TEACHER ACCREDITATION'

Additional Instructions in Scope Notes

A scope note may give additional instructions to indexers. For example, it may remind indexers of other types of terms that they should assign:

HOSPITALIZATION
SN ASSIGN ALSO TERMS FOR THE CONDITIONS FOR WHICH PATIENTS WERE HOSPITALIZED, IF APPLICABLE

Suggesting Indexers Consider More Specific Alternatives

A scope note may suggest that the term not be used if a more specific term is appropriate; for example,

EQUIPMENT
SN BROAD TERM. PREFER TERMS SPECIFYING TYPES OF EQUIPMENT IF POSSIBLE; FOR EXAMPLE, 'OFFICE EQUIPMENT'

Instructions for Synthesis

In a synthetic thesaurus, instructions for synthesis may appear in scope notes; for example,

HISTORY
SN APPEND ALSO AS A SUBDIVISION AFTER TERMS DESIGNATING DISCIPLINES, ACTIVITIES, LIVING THINGS, ETC.; FOR EXAMPLE, 'INTERCROPPING - HISTORY', 'GOATS - HISTORY'

Informativeness of Scope Notes

Information included in a scope note should be helpful to users of the thesaurus as indexers or searchers.

It should add to what the term already says by itself. Simply repeating the term or giving an obvious definition of an unambiguous term is not helpful.

Remember that a thesaurus is not a dictionary, an encyclopedia, or even an index.

Form of Scope Notes

Scope notes should be **well formed**.

They should contain **no spelling errors**.

Many scope notes do not use complete sentences. You can use noun and verb phrases instead. Nevertheless, the **syntax** should be correct.

Summary

To sum up, scope notes may

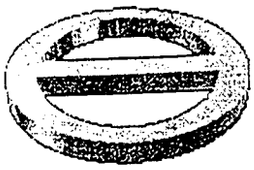
1. give definitions
2. indicate which concepts are included or excluded
3. refer to other terms
4. provide additional instructions

and they should be

1. relevant to indexing and searching
2. well-formed

Quiz on scope notes (requires JavaScript for Netscape Navigator 3.0)

Section 6 Section 8 Table of Contents Glossary



Section 8
Thesaurus Displays

For any thesaurus display, you may need to make several decisions. These decisions are likely to include

- which types of terms will have entries
- how to indicate special types of terms
- what types of links will be shown to other terms
- how many levels of linking will be shown
- how to indicate link types
- where the linked terms are placed
    - relative to the entry term
    - relative to each other

Any of these decisions will, of course, be constrained in various ways. For example, the thesaurus construction software that you use may produce only certain kinds of displays or may not permit you to store a mixture of upper and lower case.

Which Types of Terms Will Have Entries?

A thesaurus display might have entries only for preferred terms; for example,

...
EX-CONVICTS
EYE EXAMINATIONS
EYE PATCHES
EYEGLASSES
EYES
FABLES
...

At least one of the displays, however, should provide entries for non-preferred terms as well, to allow users to browse through these for the correct preferred terms:

...
EXTREMISM
EX-CONVICTS
EX-MILITARY PERSONNEL
EYE CATCHERS
EYE EXAMINATIONS
EYE PATCHES

EYEGLASSES
EYES
FABLES
FABRIC DESIGN DRAWINGS
...

One of the displays might include entries only for **top terms**, preferred terms that have no broader terms. This choice is often combined with indicating multiple levels of narrower terms, in a tree display, as discussed below.

How to Indicate Special Types of Terms

You may want to mark certain kinds of terms in special ways. For example, you might put all the non-preferred terms in italics:

...
EXTREMISM
EX-CONVICTS
EX-MILITARY PERSONNEL
EYE CATCHERS
EYE EXAMINATIONS
EYE PATCHES
EYEGLASSES
EYES
FABLES
FABRIC DESIGN DRAWINGS
...

Of course, users of the thesaurus should be able to tell that a term is a non-preferred term if it has a USE reference after it, but displaying the term differently will serve as an added reminder.

The examples used in this tutorial generally show terms in all upper case. This is to emphasize that the distinction between upper and lower case should normally not be significant in indexing and searching using a controlled vocabulary. Nevertheless, you may prefer a mixture of upper and lower case for your thesaurus displays to make them easier to read. Mixing upper and lower case may be especially helpful for longer elements such as scope notes:

GOGGLES
SN ***Protective eye coverings.***

What Types of Links Should be Shown to Other Terms

Taken as a whole, your thesaurus displays should cover all the term links that are important to the people who will use the thesaurus. In individual displays, you may choose to include only certain links. For example, a brief display might include only "USE" references:

EXTREMISM
USE RADICALISM
EX-CONVICTS
EX-MILITARY PERSONNEL
USE VETERANS
EYE CATCHERS
USE ARCHITECTURAL FOLLIES

EYE EXAMINATIONS
EYE PATCHES
EYEGLASSES
EYES
FABLES
FABRIC DESIGN DRAWINGS
USE TEXTILE DESIGN DRAWINGS

An entry in one of the displays will usually give all the links:

EYEGLASSES
UF *SPECTACLES*
BT MEDICAL EQUIPMENT AND SUPPLIES
BT OPTICAL DEVICES
NT MONOCLES
NT SUNGLASSES
RT CONTACT LENSES
RT EYE PATCHES
RT GOGGLES

Similarly, at least one of the displays would include the scope notes:

ALIDADES
SN *TELESCOPIC SITING DEVICES USED AS PART OF A SHIP'S NAVIGATIONAL EQUIPMENT FOR TAKING BEARINGS*
BT SCIENTIFIC EQUIPMENT
BT TELESCOPES
RT NAVIGATION

How Many Levels of Linking will be Shown

In one of your displays, you may wish to show indirectly linked terms as well as those linked directly to the entry term. This is mostly useful with links representing hierarchical relations. The display could indicate more than one level of broader term:

MONOCLES
BT EYEGLASSES
. BT OPTICAL DEVICES
. . BT EQUIPMENT
. BT MEDICAL EQUIPMENT AND SUPPLIES

Likewise, the display could indicate more than one level of narrower term:

OPTICAL DEVICES
NT BINOCULARS
NT CONTACT LENSES
NT EYEGLASSES
. NT MONOCLES
. NT SUNGLASSES
NT GOGGLES
...

As mentioned above, such a cascade of narrower terms is often used with top terms as entry terms.

How to Indicate Link Types

You can often omit the symbols for the different kinds of links if it is obvious what they are. So, you need to use a symbol such as "RT" only once for a series of terms all linked to the entry term with an "RT" reference:

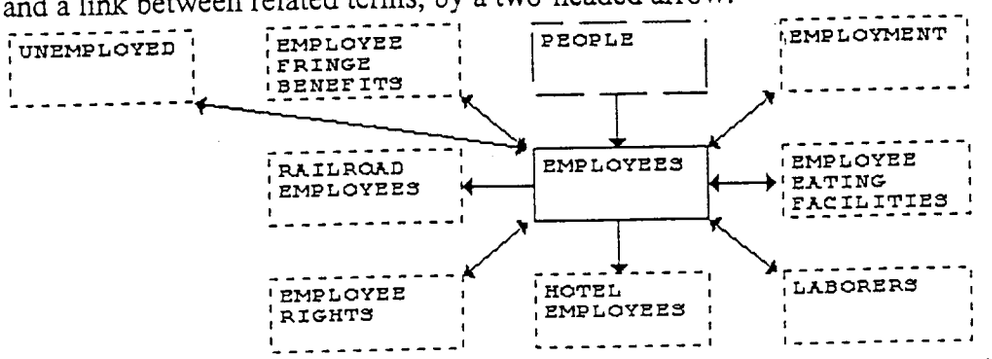
...
RT EMPLOYEE EATING FACILITIES

EMPLOYEE FRINGE BENEFITS
EMPLOYEE RIGHTS
EMPLOYMENT
LABORERS
UNEMPLOYED

Similarly, if all the links in a display are of the same kind, as in a hierarchical display, you do not need to use a distinctive symbol:

EQUIPMENT
. AIRPLANE EQUIPMENT
.. AIRPLANE PROPELLERS
.. AIRPLANE WINGS
. ANCHORS
. APPLIANCES
.. AIR CONDITIONERS
.. DISHWASHING MACHINES
.. FREEZERS
.. REFRIGERATORS
.. TOASTERS
.. VACUUM CLEANERS
...

In a graphic display, different types of links can be represented by different kinds of arrows. For example, a link from a broader term to a narrower term can be represented by a one-headed arrow; and a link between related terms, by a two-headed arrow:



Where the Linked Terms are Placed

Relative to the Entry Term

In a printed display, you will usually want the entry term to appear at the top left, because this makes it easy to search for. Variations are possible, though. For example, broader terms may be displayed

above the entry term and narrower terms below:

. MEDICAL EQUIPMENT AND SUPPLIES
. . EQUIPMENT
. OPTICAL DEVICES
EYEGLASSES
. MONOCLES
. SUNGLASSES

In an online display of a single entry, there is more flexibility. For example, the broader terms can be arrayed on the left, the narrower terms on the right, and the related terms above and below:

MEDICAL EQUIPMENT AND SUPPLIES . EQUIPMENT . . OPTICAL DEVICES .	CONTACT LENSES EYE PATCHES EYEGLASSES GOGGLES	. MONOCLES . SUNGLASSES

Relative to Each Other

If all the links are indicated in the same position relative to the entry term, the best order to follow is generally

- scope notes
- non-preferred equivalent terms
- broader terms
- narrower terms
- related terms

For example,

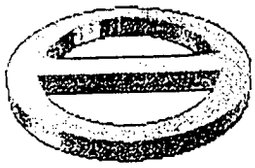
EMPLOYEES
SN *PERSONS IDENTIFIED AS WORKING FOR ANOTHER, BUT WHERE THE NATURE OF THE OCCUPATION, BUSINESS, OR INDUSTRY IS NOT KNOWN.*
UF *PERSONNEL*
STAFF
WORKERS
BT PEOPLE
NT HOTEL EMPLOYEES
RAILROAD EMPLOYEES
RT EMPLOYEE EATING FACILITIES
EMPLOYEE FRINGE BENEFITS
EMPLOYEE RIGHTS
EMPLOYMENT
LABORERS
UNEMPLOYED

Within a group of terms linked in the same way to the entry term, the order is most commonly alphabetical, as in the example just given. Sometimes, however, you may wish to adopt a systematic order by subcategorizing the link types, especially if the lists are very long.

Section 7 Table of Contents Glossary

Last updated August 2, 2002.
This page maintained by Prof. Tim Craven
E-mail (text/plain only): craven@uwo.ca
Faculty of Information and Media Studies
University of Western Ontario,
London, Ontario
Canada, N6A 5B7

Copyright © 1997 The University of Western Ontario



Glossary

Broader term
The superordinate term in a hierarchical relation.

BT
A symbol used in a thesaurus to identify the following terms as broader terms to the heading term.

Chain
A sequence of terms in which the class represented by each term includes all the classes represented by the terms that follow that term.

Controlled indexing
Indexing with terms from a controlled vocabulary, such as a thesaurus.

Enumerative
Of an indexing or classification scheme, listing or enumerating terms explicitly, rather than making provision for synthesizing them.

Equivalent term
A term in a controlled vocabulary, such as a thesaurus, that is treated as if it means the same thing as another term.

Extraction indexing
Indexing with terms from the text or title of the item indexed.

Hierarchical relation
A semantic relation in which one term is strictly subordinate to the other; for example, a genus/species relation.

Homograph
A term with the same spelling as, but a different meaning from, another term.

Narrower term
The subordinate term in a hierarchical relation

Non-preferred term
A term in a controlled vocabulary, such as a thesaurus, that may not be used as a controlled indexing term, but may be looked up.

NT
A symbol used in a thesaurus to identify the following terms as narrower terms to the heading term.

Postcoordination
The combination of terms at the time of retrieval to form a compound search specification that corresponds to no single index term used in indexing.

Precision ratio
The ratio of the number of relevant items retrieved to the total number of items retrieved.

Precoordination
The combination of terms or other components to form compound terms at or before the time of indexing.

Preferred term
A term in a controlled vocabulary, such as a thesaurus that may be used as a controlled indexing term.

Quasi-synonym
An equivalent term that is not a synonym.

Recall ratio
The ratio of the number of relevant items retrieved to the total number of relevant items

indexed.

Related term
A term in a semantic relation, but not in a hierarchical relation, to another term.

Relevant
Appropriately retrieved in response to a given query.

RT
A symbol used in a thesaurus to indicate that the following terms are related terms to the heading term.

Scope note
A note attached to a term in a controlled vocabulary, such as a thesaurus, that gives guidance on how to use the term.

Semantic relation
A relation between terms that is true as a matter of general knowledge, rather than depending on what the terms refer to in some particular document.

SN
A symbol used in a thesaurus to indicate that what follows is a scope note to the heading term.

Syntactic relation
A relation between terms that depends on what the terms refer to in some particular document, rather than being true as a matter of general knowledge.

Synthetic
Of an indexing or classification scheme, making provision for the synthesis of terms out of components, rather than listing or enumerating them explicitly.

Thesaurus
A device for vocabulary control, usually for a specific subject area, indicating preferred terms, non-preferred terms, and semantic relations between terms; the terms are in ordinary human language.

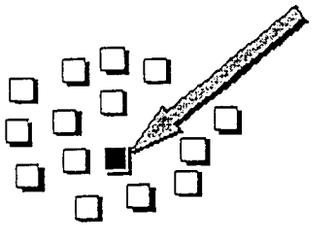
UF
A symbol used in a thesaurus to indicate that the following term or terms are not used in indexing and that the heading term is used instead.

USE
A symbol used in a thesaurus to indicate that the heading term is not used in indexing and that the following term or terms are used instead.

Introduction

Last updated February 19, 2002.
This page maintained by Prof. Tim Craven
E-mail (text/plain only): craven@uwo.ca
Faculty of Information and Media Studies
University of Western Ontario,
London, Ontario
Canada, N6A 5B7

Copyright © 1997 The University of Western Ontario



Willpower Information
Information Management Consultants

Thesaurus principles and practice

*This paper was originally presented at a workshop "Thesauri for museum documentation" held at the Science Museum, London, on 24th February 1992. The proceedings of the workshop have been published by the **mda** (formerly the Museum Documentation Association).*

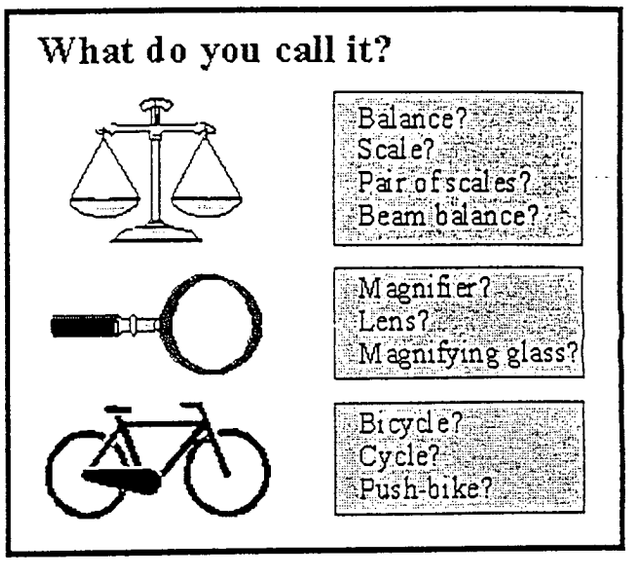
Contents

1. Why do we need a thesaurus?
2. A limited list of indexing terms
3. Hierarchical relationships
4. Related terms
5. Definitions and scope notes
6. Form of the thesaurus
7. Special factors relating to museum objects
    - Abstract terms and disciplines
    - Singular or plural terms
    - Parts and wholes
    - Polyhierarchies
8. Use of a thesaurus when cataloguing
9. Use and modification of existing thesauri
10. Thesaurus maintenance
11. What sort of fields is a thesaurus appropriate for?
12. Other subject retrieval techniques
    - Classification schemes
    - Free text

- Bibliography on thesaurus construction and use

1 Why do we need a thesaurus?

One of the reasons for documenting our collections is that we wish to be able to find objects of a particular kind. We may ask "What thermometers do we have in the collection?", "What arrowheads?", "What frocks?", "What whales?" or "What textile machinery?"



The simple answer is that we give each item a "name", and then we can create a file of index cards, or a computer file, in which we can search for these names and expect to find all the appropriate items. This is the concept of the *simple name* field in the MDA data structure. It is straightforward at first, and seems intuitive, but once you have documentation which has been built up over time, perhaps by many different people, problems creep in unless there are rules and guidelines to maintain consistency.

The word *thesaurus* is a rather fancy name, which has acquired a certain mystique, because it is often bandied about as something necessary for effective information retrieval, but something which sounds as though it will involve a lot of work. I have often heard curators say "That's all very well if you have the time and resources, but I have this great backlog of cataloguing to do, and I would never get through the half of it if I had to spend time setting up anything as complicated as a thesaurus. What I need is a simple list of names which I can use to index my objects."

My main purpose in this paper is to make three points:

- A simple name list without some rules will rapidly become a mess.
- Only three simple rules are needed; using them will make life easier for you, not harder.
- So long as you stick to these rules, you can take an existing thesaurus and adapt it to your needs; you are not limited to using the terms which are listed in it already, and you are not obliged to use more detail than you need.

What are these rules?

1. **Use a limited list of indexing terms, but plenty of entry terms**
   -- link these with `USE` and `USE FOR (UF)` relationships.
2. **Structure terms of the same type into hierarchies**
   -- link these with `BROADER TERM/NARROWER TERM (BT/NT)` relationships.
3. **Remind users of other terms to consider**
   -- link these with `RELATED TERM/RELATED TERM (RT/RT)` relationships.

I shall consider each of these rules in turn.

Contents list Home page

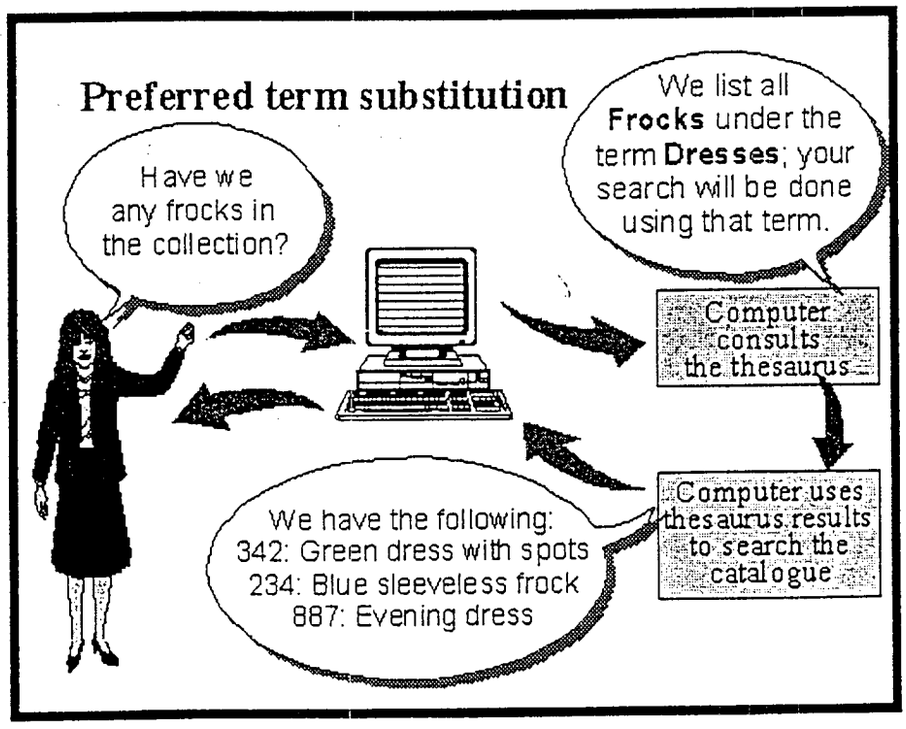
2 A limited list of indexing terms

A major purpose of a thesaurus is to match the terms brought to the system by an enquirer with the terms used by the indexer. Whenever there are alternative names for a type of item, we have to choose one to use for indexing, and provide an entry under each of the others saying what the preferred term is. If we index all full-length ladies' garments as *dresses*, then someone who searches for *frocks* must be told that they should look for *dresses* instead.

This is no problem if the two words are really synonyms, and even if they do differ slightly in meaning it may still be preferable to choose one and index everything under that. I do not know the difference between *dresses* and *frocks* but I am fairly sure that someone searching a modern clothing collection who was interested in the one would also want to see what had been indexed under the other. We normally do this by linking the terms with the terms USE and USE FOR , thus:

Dresses USE FOR *Frocks*
Frocks USE *Dresses*

This may be shown in a printed list, or it may be held in a computer system, which can make the substitution automatically. If an indexer assigns the term Frocks, the computer will change it to Dresses, and if someone searches for Frocks the computer will search for Dresses instead, so that the same items will be retrieved whichever term is used. A friendly computer will explain what it is doing, so that the user is not puzzled by being given items with terms different from those asked for.



USE and USE FOR relationships are thus used between synonyms or pairs of terms which are so nearly the same that they do not need to be distinguished in the context of a particular collection. Other examples might be:

Cloaks	USE	*Capes*
Capes	USE FOR	*Cloaks*
Nuclear energy	USE	*Nuclear power*
Nuclear power	USE FOR	*Nuclear energy*
Baby carriages	USE	*Perambulators*
Perambulators	USE FOR	*Baby carriages*
Perambulators	USE FOR	*Prams*
Prams	USE	*Perambulators*

If we name objects, we want to be as specific as possible. If we have worked hard to discern subtle distinctions in nature, type or style, we certainly want to record these. The point is that the thesaurus is not the place to do this. Detailed *description* of an object is the job of the catalogue record; the job of the thesaurus, and the index which is built by allocating thesaurus terms to objects, is to provide useful *access points* by which that record can be retrieved.

USE and USE FOR relationships can also be used to group similar items together, because too much specificity is as bad as too little. If we have a small clothing collection, containing ten jackets, it is more useful to give them all the index term *jackets* than to create many specific categories. Anyone searching our catalogue will then be able to search on the single term *jackets* and see a list of the ten items, each with a description of exactly what kind of jacket it is, as follows:

Jackets:

1. Anorak in green cotton, England, 1985.
2. Tweed sports jacket, Hawick, Scotland
3. Silk bolero with floral embroidery, Spanish, 1930.

If we used all the possible specific names, each of which would have only one or two items in it, such as *blazers, dinner jackets, boleros, donkey jackets, anoraks, flying jackets, sports jackets*, and so on, enquirers would have to search the catalogue under each name in turn in order to find all the jackets in the collection, and they would never be sure that there was not a kind of jacket that they had overlooked.

To help enquirers who approach the system by one of these terms, we therefore create the references:

Blazers	USE	*Jackets*
Dinner jackets	USE	*Jackets*

and so on.

Contents list Home page

3 Hierarchical relationships

If we have a hundred jackets, a list under a single term will be too long to look through easily, and we should use the more specific terms. In that case, we have to make sure that a user will know what terms there are. We do this by writing a list of them under the general heading, thus:

Jackets
NT *Anoraks*
Blazers
Boleros
Dinner jackets
Donkey jackets

Flying jackets
Kagouls
Sports jackets

We could just invert terms and rely on the alphabet to bring them together, in a list such as

Jackets, dinner
Jackets, donkey

Jackets, flying
Jackets, sports

but this is unreliable and subject to the vagaries of the language, which does not always describe a specific type of item by an adjective preceding the generic name. We have to accommodate types of jacket which have their own distinctive names such as *Anoraks* or *Blazers*.

In both the above cases, it is important that the terms which are linked are of the same type. That is to say that any narrower term must be a specific case of the broader term, and able to inherit its characteristics. (The developers of Object Oriented Programming have recently discovered this idea, which has been known to the worlds of information science and biological taxonomy for a very long time.) Thus if we say that *Blazers* is a narrower term of *Jackets*, we mean that every blazer is, whatever else it may be, inherently a jacket, and that it has the characteristics which define a jacket.

Mice can properly be said to be a narrower term of *Rodents*, because all mice are inherently rodents, but it is not correct to list *Mice* as a narrower term of *Pests*, because some mice, such as laboratory mice and pet mice, are not pests. The idea is to have relationships in the thesaurus which are always true,

Broader and narrower terms
Hierarchical relationships

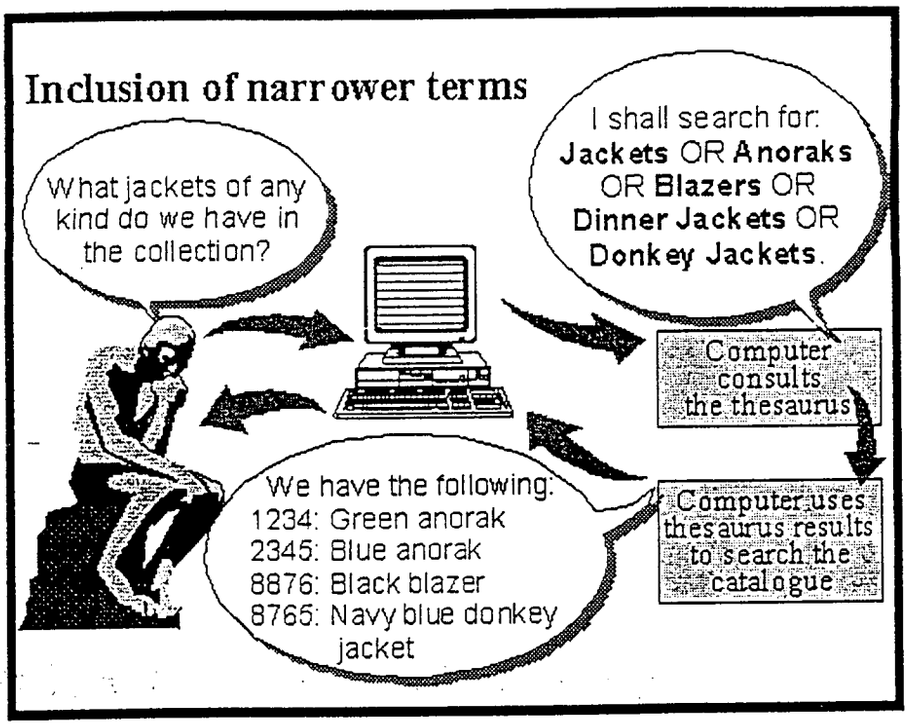
- **Relationships must be independent of context**
- **Terms must represent the same type of entity**

✓	*Mice*	BT	*Rodents*
	Rodents	NT	*Mice*
✓	*Shoes*	BT	*Footwear*
	Footwear	NT	*Shoes*
✗	*Mice*	BT	*Pests*
	Pests	NT	*Mice*

irrespective of context. In the same way, it would not be correct to list *Buses* as a narrower term of *Diesel-engined vehicles*, although many of them are; if we have a diesel-engined bus in our collection, we should show this by giving it the two terms *Buses* and *Diesel-engined vehicles.*

	Shoes	BT	*Shoemaking*
	Shoemaking	NT	*Shoes*

Good computer software should allow you to search for "*Jackets and all its narrower terms*" as a single operation, so that it will not be necessary to type in all the possibilities if you want to do a generic search:



Contents list Home page

4 Related terms

If we restrict the hierarchical relationship to true specific/generic relationships, we need another mechanism to draw attention to other terms which an indexer and a searcher should consider. These are RELATED TERMS of the starting term. Related terms may be of several kinds:

1. Objects and the discipline in which they are studied, such as Animals and Zoology.
2. Process and their products, such as Weaving and Cloth.
3. Tools and the processes in which they are used, such as Paint brushes and Painting.

It is also possible to use the RELATED TERM relationship between terms which are of the same kind, not hierarchically related, but where someone looking for one ought also to consider searching under the other, e.g. *Beds* RT *Bedding*; *Quilts* RT *Feathers*; *Floors* RT *Floor coverings*.

Contents list Home page

5 Definitions and scope notes

A thesaurus is not a dictionary, and it does not normally contain authoritative definitions of the terms which it lists. It could perfectly well do this, but a lot more work would be required to develop it in this way. In an automated system, however, the thesaurus would be a logical place to record information which is common to all objects to which a term might be applied, for example notes on the history and origin of Anoraks or the identifying characteristics and lifestyle of Mice (or perhaps *Mus musculus* in a taxonomic thesaurus).

Where there is any doubt about the meaning of a term, or the types of objects which it is to represent, a SCOPE NOTE (SN) is attached to it. For example,

Fruit
SN *distinguish from Fruits as an anatomical term*
BT *Foods*

Preserves
SN *includes jams*

Neonates
SN *covers children up to the age of about 4 weeks; includes premature infants*

Table 1: Samp thesaurus - hierarchical sequence

knitwear
> cardigans
> pullovers
outerwear
> blouses
> cardigans
> coats
> > raincoats
> dresses
> jackets
> > anoraks
> > blazers
> > dinner jacke
> > donkey jack
> > reefer jacke
> leggings
> pullovers
> rainwear
> > raincoats
> shawls
> shirts
> skirts
> suits
> trousers
> > jeans
> > shorts
> > slacks

6 Form of the thesaurus

A list based on these relationships can be arranged in various ways; alphabetical and hierarchical sequences are usually required, and thesaurus software is generally designed to give both forms of output from a single input. A typical simple thesaurus of a few clothing terms is shown in Tables 1 and 2.

Table 2: Sample thesaurus - alphabetical sequence

anoraks
BT jackets

blazers
BT jackets

blouses
UF smocks
BT outerwear

breeches
USE trousers

capes
USE coats

cardigans
SN knitted jackets with front opening
BT knitwear
outerwear

cloaks
USE coats

coats
UF capes
cloaks
overcoats
BT outerwear
NT raincoats

dinner jackets
BT jackets

donkey jackets
BT jackets

dresses
UF frocks
BT outerwear

duffel jackets
USE reefer jackets

frocks
USE dresses

jackets
BT outerwear
NT anoraks
blazers
dinner jackets
donkey jackets
reefer jackets

jeans
BT trousers

jumpers
USE pullovers

knitwear
NT cardigans
pullovers

leggings
BT outerwear

outerwear
NT blouses
cardigans
coats
dresses
jackets
leggings
pullovers
rainwear
shawls
shirts
skirts
suits
trousers

overcoats
USE coats

pullovers
UF jumpers
sweaters
BT knitwear
outerwear

raincoats
BT coats
rainwear

rainwear
BT outerwear
NT raincoats

reefer jackets
UF duffel jackets
BT jackets

shawls
UF wraps (clothing)
BT outerwear

shirts
BT outerwear

shorts
BT trousers

skirts
BT outerwear

slacks
BT trousers

smocks
USE blouses

suits
BT outerwear

sweaters
USE pullovers

trousers
UF breeches
BT outerwear
NT jeans
shorts
slacks

wraps (clothing)
USE shawls

Contents list Home page

7 Special factors relating to museum objects

7.1 Abstract terms and disciplines

Many thesauri have been created with the intention of being used to index documentary material, and thus they include many terms which relate to abstract concepts, disciplines and areas of discussion, as well as the names of concrete objects which are of primary interest to museums. We have to be careful to be consistent in how we use these terms. The most straightforward way is to concentrate first on what objects actually are - spades are *Spades* and should be given this term, rather than the area in which they are used, whether it is gardening or gravedigging.

You may well wish to allocate abstract and discipline terms to objects too, so that you can retrieve all the objects to do with *Dentistry, Laundry, Warfare* or *Food preparation.* These terms can also be included in the thesaurus, so long as they are not given hierarchical relationships to names of objects. They should be given RT relationships to an appropriate level of object terms.

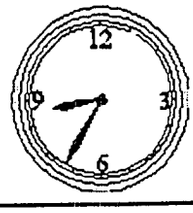
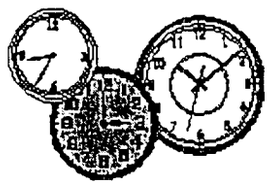
Some thesauri, such as ROOT, interfile terms of different types in their hierarchical display. Indentation in such cases does not necessarily indicate a BT/NT relationship. The relationships are shown in ROOT's alphabetical sequence, and it is unfortunate that they are not distinguished in the hierarchical one.

Because these abstract terms do not describe what the object *is*, they could be put into a field in the catalogue record labelled *concept* or *subject*, distinct from the field containing terms which *name* the object. I do not think that such a distinction will generally be helpful to users, however, and there seems to be no disadvantage in putting both types of term into a single field so that they can easily be searched as alternatives or in combination. Such a field would not be correctly called *name* and I therefore prefer to call it simply *indexing terms* or *subject indexing terms.*

7.2 Singular or plural terms

There has been much discussion on whether thesaurus terms should be expressed in the singular or the plural. I believe that the difficulty arises from different views of what is being done when a term is assigned to an object record. If a cataloguer thinks that (s)he is naming the object in hand, (s)he will naturally use the singular: "This is a clock". If (s)he is assigning the object to a category of similar objects, the thought will be "This belongs in the category of clocks". An enquirer will normally ask for a category, so the latter form will be more natural and logical.

The point is not a trivial one, because as discussed in section 2 above there is a conceptual difference between naming or describing an object and grouping it with others so that it can be found. Both are essential steps, but an information retrieval thesaurus is primarily concerned with grouping.

Singular or plural terms?	
The cataloguer thinks: "This is a *clock*".	
The enquirer asks: "What *clocks* do you have?"	
Prefer plural terms because:	

- We should design the catalogue to fit the way the user thinks.
- *Clocks* is the name of a category, including many types, so plural is more logical.

The British Standard for thesaurus construction recommends that plural terms should be used, except for a few well-defined cases, and my view is that this practice should be followed. Unfortunately, there are many records in museum collections which have been given singular "object names", and the work of changing these to plurals in a move to a thesaurus structure may be so great as to require some compromise.

7.3 Parts and wholes

The British Standard recommends that when indexing parts or components, separate terms should be assigned for the component and for the object of which it forms part, so that aircraft engines would be indexed by the two terms *Aircraft* and *Engines*. This causes problems in a museum collection, however, because items indexed in this way would be retrieved in a search for *Aircraft*, when only whole aircraft were being sought. It therefore seems preferable to use a term such as *Aircraft components*. A particular engine may well be an aircraft component, but it is not an aircraft. Similarly a timer from a cooker can be indexed by the terms *Timers* and *Cooker components*, and a handle broken from a vase might be indexed as *Handles* and *Vase fragments*. There needs to be local agreement on how this approach is to be applied to a particular collection.

In the thesaurus, `BT/NT` relationships can be used for parts and wholes in only four special cases: parts of the body, places, disciplines and hierarchical social structures.

7.4 Polyhierarchies

As shown in the sample thesaurus above, a term can have several broader terms, if it belongs to several broader categories. The thesaurus is then said to be polyhierarchical. *Cardigans*, for example, are simultaneously *Knitwear* and *Jackets*, and should be retrieved whenever either of these categories is being searched for.

Art and architecture thesaurus (AAT), regrettably, allows a term to have only one broader term, and I think that this is a serious drawback to its usefulness as an information retrieval tool. It would take more space to repeat full hierarchies under each of several broader terms in a printed version, but this can be overcome by using references, as ROOT does. There is no difficulty in displaying polyhierarchies in a computerised version of a thesaurus.

Contents list Home page

8 Use of a thesaurus when cataloguing

A thesaurus is an essential tool which must be at hand when indexing a collection of objects, whether by writing catalogue cards by hand or by entering details directly into a computer. The general principles to be followed are:

1. Consider whether a searcher will be able to retrieve the item by a combination of the terms you allocate.
2. Use as many terms as are needed to provide required access points.
3. If you allocate a specific term, do not also allocate that term's broader terms.
4. Make sure that you include terms to express what the object is, irrespective of what it might have been used for.

If you have a computerised thesaurus, with good software, this can give you a lot of direct help. Ideally it should provide pop-up windows displaying thesaurus terms which the cataloguer can choose from and then "paste" directly into the catalogue record without re-typing. It should be possible to browse around the thesaurus, following its chain of relationships or displaying tree structures, without having to exit the current catalogue record, and non-preferred terms should automatically be replaced by their preferred equivalents. A cataloguer should be able to "force" new terms onto the thesaurus, flagged for review later by the thesaurus editor. When editing thesaurus relationships, reciprocals should be maintained automatically, and it should not be possible to create inconsistent structures.

---

9 Use and modification of existing thesauri

As there are many thesauri in existence already, it is worth considering seriously whether one of these can be used before embarking on the job of creating a new one for a particular museum or collection. So long as the general principles are followed, you should be able to expand a thesaurus to give you more detail if you need it, or truncate some sections at a high level if they contain more detail than your collections justify. So long as the relationships are universally true, it should be possible to combine sections of thesauri developed by different museums and thus avoid duplication of work.

Even when using an authoritative thesaurus, some care is needed, and I have mentioned some limitations of ROOT and AAT in 7.1 and 7.4 above. It is still much easier to base your work on something like these than to build your own from scratch, unless you have a very specialised collection.

---

10 Thesaurus maintenance

Someone has to be responsible for this. New terms can be suggested, and temporarily "forced" into the thesaurus by cataloguers as they catalogue objects, but someone has to review these terms regularly and either accept them and build them into the thesaurus structure, or else decide that they are not appropriate for use as indexing terms. In that case they should generally be retained as non-preferred terms with USE references to the preferred terms, so that people who seek them will not be frustrated. An encouraging thought is that once the initial work of setting up the thesaurus has been done, the number of new terms to be assessed each week should decrease, and many systems have operated successfully in the past with printed thesauri, which are quite difficult to keep up to date.

Contents list

Home page

11 What sort of fields is a thesaurus appropriate for?

A thesaurus is not a panacea which will meet all subject retrieval needs. It is particularly appropriate for fields which have a hierarchical structure, such as names of objects, subjects, places, materials and disciplines, and it might also be used for styles and periods. A thesaurus proper would not normally be used for names of people and organisations, but a similar tool, called an authority file is usually used for these. The difference is that while an authority file has preferred and non-preferred relationships, it does not have hierarchies.

[Authority files and thesauri are two examples of a generalised data structure which can allow the indication of any type of relationship between two entries, and modern computer software should allow different types of relationship to be included if needed.]

12 Other subject retrieval techniques

A thesaurus is an essential component for reliable information retrieval, but it can usefully be complemented by two other types of subject retrieval mechanism.

12.1 Classification schemes.

While a thesaurus inherently contains a classification of terms in its hierarchical relationships, it is intended for specific retrieval, and it is often useful to have another way of grouping objects. This may relate to administrative distribution of responsibility for "collections" within a museum, or to subdivisions of these collections into groups which depend on local emphasis. It is also often necessary to be able to print a list of objects arranged by subject in a way which differs from the alphabetical order of thesaurus terms. Each subject group may be expressed as a compound phrase, and given a classification number or code to make sorting possible.

12.2 Free text.

It is highly desirable to be able to search for specific words or phrases which occur in object descriptions. These may identify individual items by unique words such as trade names which do not occur often enough to justify inclusion in the thesaurus. A computer system may "invert" some or all fields of the record, i.e. making all the words in them available for searching through a free-text index, or it may be possible to scan records by reading them sequentially while looking for particular words. The latter process is fairly slow, but is a useful way of refining a search once an initial group has been selected by using thesaurus terms.

This document is at http://www.willpowerinfo.co.uk/thesprin.htm
Revised 13th February 1998

Comments and feedback on content or presentation are welcome and should be sent to Leonard Will at L.Will@willpowerinfo.co.uk

Copyright © Willpower Information, 1998.

Contents list Home page

Theoretical Foundations of Thesaurus-Construction and Some Methodological Con...
CHAI KIM
Journal of the American Society for Information Science (pre-1986); Mar/Apr 1973; 24, 2; ABI/INFORM Global
pg. 148

Theoretical Foundations of Thesaurus-Construction and Some Methodological Considerations for Thesaurus-Updating

It was argued that the present-day thesaurus-construction and maintenance rules and conventions are not theoretically based. For this reason, there are few rules and conventions for updating a thesaurus. Consequently, most of the thesauri adopted by operating information storage and retrieval systems are not systematically updated. In order to investigate how thesauri are actually updated, a survey was conducted. The working hypothesis was that the communication process between authors and readers is linear in nature ("one-way" communication allowing no reciprocal feedback) if a thesaurus utilized in a system is not updated by *both* indexers and question negotiators. Findings show that thesauri viewed from the communications point of view do not allow a cybernetic process of communication ("both-way" communication). The survey indicated that the present practice of updating thesauri is largely done by indexers alone. No attempt was made to develop a theory of thesaurus-construction and updating. It was, however, argued that such a theory, if developed, should at least account for the concepts of meaning and knowledge. Within this theoretical framework, two techniques are suggested to be considered for the systematic updating of a thesaurus.

CHAI KIM

Interdisciplinary Doctoral Program in Information Science
University of Pittsburgh

• Introduction

There are three general types of tools designed to control vocabularies of a given information storage and retrieval system—classification schemes, subject heading authority lists, and thesauri. This paper is primarily on thesauri as they are used in information storage and retrieval systems.

Three specific goals of this paper are: 1) to investigate whether or not there are theoretical foundations upon which thesaurus-construction and maintenance rules, conventions and guidelines are based; 2) to examine the state-of-the-art of the thesaurus-updating; and finally, 3) to suggest possible research methods to be utilized for systematically updating thesauri.

The major argument of this paper is that because of our poor (or lack of) understanding of the underlying theoretical foundations upon which a thesaurus is built, we are without any systematic rules, conventions or guidelines for updating a thesaurus. Hence we continue to build *new* thesauri when *old* ones cease to be useful.

We are inundated with a huge quantity of literature on *how* a thesaurus is built, but very little—or no—literature which discusses *why* such and such rules need to exist. We have hundreds of articles and numerous "rules and conventions" on how a thesaurus is built but merely a few paragraphs on how a thesaurus should be updated once it is built.

• Definitions and Functions

Let us start by examining the statements regarding definitions and functions of thesauri. To do this, the following thesaurus rules and conventions or guidelines were examined. They are: 1) Engineers Joint Council. "Thesaurus Rules and Conventions" (in) *Thesaurus of Engineering and Scientific Terms:* a list of engineering and related scientific terms and their relationships for use as vocabulary reference in indexing and retrieving technical information. New York: Engineers Joint Council, 1967; 2) COSATI. *Guidelines for the Develop-*

Reproduced with permission of the copyright owner. Further reproduction prohibited without permission.

ment of Information Retrieval Thesauri. Washington, D.C.: Govt. Print. Office, 1967; 3) U.S. Office of Education. ERIC. *Rules for Thesaurus Preparation.* Washington, D.C.: U.S. Dept. of Health, Education, and Welfare. Bureau of Research, 1969; 4) UNESCO. *Guidelines for the Establishment and Development of Monolingual Scientific and Technical Thesauri for Information Retrieval.* Paris, 1970.

(Throughout this paper, for convenience, those cited above will be referred to respectively as: 1) EJC *1967-Rules;* 2) COSATI *1967-Guidelines;* 3) ERIC *1969-Rules;* and 4) UNESCO *1970-Guidelines.*)

Definitions

According to the UNESCO *1970-Guidelines*, a thesaurus is defined as "a controlled and dynamic vocabulary of semantically and generically related terms which comprehensively covers a specific domain of knowledge. This vocabulary is a systematical and/or alphabetical collection of descriptors, non-descriptors (auxiliary terms) as well as indicators of their relationships (*1*)."

According to the COSATI *1967-Guidelines*, "a thesaurus is a compilation of selected terms with appropriate term interrelationships displayed in such a way as to promote maximum consistency in the description of concepts for indexing and searching (*2*)."

In the case of our first citation, one of the important questions we must ask is about the semantic-theoretical foundation upon which "semantically and generically related terms which . . . covers a specific domain of knowledge" is based. In other words, there must be some semantic theory which accounts for such relationships between terms. The same question can be asked regarding the COSATI *1967-Guidelines* on the "appropriate term interrelationships." When terms are said to be appropriately interrelated, upon which theories are they based, i.e., there must be a theory that accounts for such interrelationships.

In short, both Guidelines we have cited above fail to mention anything about the underlying or assumed semantic theory upon which statements on construction of a thesaurus are based. By and large, rules and conventions, or guidelines, for thesaurus-construction tell us *how* to establish term-interrelationships by utilizing such schemes as "Broader Term," "Narrower Term," "Related Term," etc., but fail to, or choose not to, discuss the theoretical foundations underlying such rules and conventions or guidelines. This is perfectly understandable —after all they are the books of rules. Judging from the literature search, however, it is not at all understandable why so few people are concerned with theoretical problems of thesaurus-construction and maintenance.

Because of this lack of theoretical foundations, we are not altogether sure as to what is meant when we read, say, the definition of the UNESCO *1970-Guidelines:* "a controlled and dynamic vocabulary of semantically and generically related terms which comprehensively covers a specific domain of knowledge." What we must ask is: What do they mean by "dynamic?"—vocabularies? or a "specific domain of knowledge?" or both? Their definition seems to indicate that only the vocabulary is dynamic. If so, are we assuming that "a specific domain of knowledge" is *always static,* but only the words with which we represent or describe it are dynamic? If we were to agree on this, upon which "theory of knowledge" is our agreement based?

It is therefore clear that a thesaurus must be founded on *at least* two different theories—a semantic theory and an epistemic theory. We concluded this by examining parts of the definitions given in the UNESCO *1970-Guidelines* and the COSATI *1967-Guidelines.* For example, from "semantically and generically related terms" and "appropriate term interrelationships," we have concluded that one cannot semantically or generically interrelate two terms unless he presupposes some theory of meaning. From "a controlled and dynamic vocabulary . . . comprehensively covers a specific domain of knowledge," we have concluded that one cannot talk about the state of knowledge—either static or dynamic—unless we assume a definition of knowledge. One more problem here that we have not discussed so far is the relationship between vocabularies and knowledge or concepts. (We will discuss this point later on.)

Functions

Kent offers some useful insight on what a thesaurus is (or ought to be):

> . . . instead of thinking of a thesaurus in its definitional sense, that is, a dictionary of synonyms and antonyms, it may be more useful to think of it as a compilation which contains the terms of a given information-retrieval system's vocabulary, arranged in some meaningful form, and which provides information relating to each term that will *enable the user of the information file to predict the relevance of responses to questions when this particular vocabulary control mechanism is used.* (Italics are mine, not the author's (*3*).)

What is explicit in this very functionally oriented definition is that a thesaurus is used to regulate both the input and output of a system.

Using the NASA thesaurus, Kent conducted an experiment to empirically determine if a thesaurus functions as a regulatory device for the output. In other words, he was interested in testing how accurately an information retrieval thesaurus "enables the user . . . to predict the relevance of responses to questions . . ." That is: when we use a thesaurus, just how much regulating power do we really acquire in controlling the output? The results of the experiment were mainly negative. Thus, Kent concluded that:

> the lack of effectiveness in regulating the output was observed to be true of the pre-thesaurally indexed NASA file, but may also apply to the file as indexed post-thesaurally . . . If it is indeed true that the post-

Reproduced with permission of the copyright owner. Further reproduction prohibited without permission.

thesaural indexing is not significantly different from the SAL indexing which was shown to be lacking in the necessary redundancy for fine control of the retrieval (output) procedures, then the only effective function being performed by the *Thesaurus* is the minor function of an authority list for the selection of descriptors (input). The major function as a tool for regulating the output of large mechanized retrieval systems remains unusable. And as a consequence, recall and precision values are likely to remain low (4).

As we can see, he went ahead and stated the twofold functions of a thesaurus: the major function is (or ought to be) as a tool for regulating the output; the minor function is as an authority list for the input. One might argue with his choice of words—"major" and "minor." In any event, it seems reasonable to expect from the usage of a thesaurus *at least* those two functions —the regulatory functions on both the input and output stages, whichever function we might call, "major" or "minor."

Let us now discuss the theoretical foundations of thesauri.

• Theoretical Foundations of Thesauri

The purpose of this section is to discuss two of the most important theories of meaning to find out how each theory defines meaning, so that we may examine the definition of meaning a thesaurus seems to presuppose.

Broadly speaking, there are two major theories of meaning: 1) Analytical (referential) theory of meaning; and 2) Operational (contextual) theory of meaning.

First, let us briefly examine the analytical theory of meaning. The best known proponents of this theory are Ogden and Richards. They explain meaning through their basic triangle model as shown in Figure 1.

Ogden and Richards view meaning as having three different components. Meaning is said to exist between relationships of three components. The relationship between *A and B* is that of *symbolization*, while the relationships between *A and C* and *B and C* are that of *reference*. Stated differently, the relationship between *A and B* is linguistic, whereas the relationships between *A and C* and *B and C* are non-linguistic. The relation between *B and C* is represented by the dotted line because Ogden and Richards felt that there is *no direct interrelationship* between words and the things they stand for. (After all, don't we say that words are *labels* of things they refer to?)

The crucial question for us to ask seems to be: Of the three different relationships—between *A and B*, *A and C*, and *B and C*—which one concerns us most? Since our business is to design and maintain a thesaurus using the natural language, it seems that we are not directly concerned with the non-linguistic relationships between *A and C* and *B and C*. Hence, they lie beyond our interests. What we are primarily concerned with is

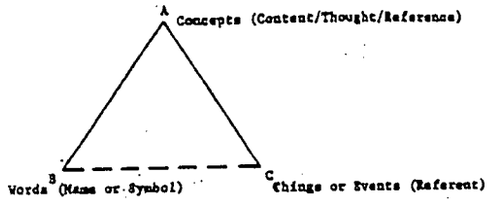


FIG. 1

the relationship between *A and B*. This is the relationship between concepts and words. Stated differently, it is the relationship between concepts and the overt expression (of them) such as books and articles.

Ullmann's definition of meaning is particularly useful for our purpose:

> It is at this point that the Ogden-Richards model does not go far enough. It gives an account of how the word acts on the hearer, but seems to neglect the speaker's point of view. For the hearer, the sequence of events will be as shown in the basic triangle: hearing the word, say, door, he will think of a door and thus understand what the speaker was saying. For the speaker, the sequence will be just the reverse: he will, for some reason or another, think of a door, and this will make him pronounce the word. There is therefore a reciprocal and reversible relationship between name and sense: if one hears the word, one will think of the thing, one will say the word. It is this reciprocal and reversible relationship between sound and sense which I propose to call the 'meaning' of the word (5).

In our context, then, meaning is the reciprocal and reversible relationship between *concepts* ("sense" in Ullmann's terminology; "thought" or "reference" in Ogden-Richards' terminology) and *words* (we may call it "overt expression"; or "symbol"—Ogden and Richards' terminology; or "name"—Ullmann's terminology; or we may even want to use "documents").

Consider the diagram below. What we are trying to do in designing a thesaurus is to explicate the relationship between concepts (1) and words (2) and at the same time we are trying to characterize words (2) by selectively displaying semantically related terms so that the use of a thesaurus (3) will result in the retrieval of not only (2) but hopefully (1) also. This continuum between (1)-(2)-(3) must always be in our mind when we design a thesaurus.

To design a thesaurus is to aim at: a) explicating meaning—the relationship between (1) and (2) in Figure 2; and b) displaying words in such a way that (2) may be retrieved when a thesaurus is used. Ideally a thesaurus, when used, should "guarantee" more than simple matching and retrieving. (It is more than the "white pages" of a telephone directory.) A thesaurus

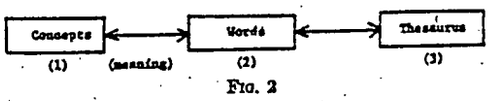


FIG. 2

Reproduced with permission of the copyright owner. Further reproduction prohibited without permission.

should not be defined, therefore, as simply a tool used for vocabulary control for input and output. We are interested in controlling vocabularies not simply to retrieve words, but primarily to retrieve words that reflect concepts. The problems of retrieving words are largely clerical in nature. The problems of retrieving words that reflect concepts add another dimension to the problem. Herein lies our challenge.

Now let us examine the second of the two theories of meaning—operational (contextual) theory of meaning.

In 1932, P. W. Bridgman in his book *The Logic of Modern Physics* emphasized a new concept of "concept" by saying: "We mean by any concept nothing more than a set of operations; the concept is synonymous with the corresponding set of operations." This concept of "operationalism" was extended to the study of meaning. In other words, concepts not subject to operations are held to be meaningless. For instance, Stuart Chase quotes Bridgman: "The true meaning of a word is to be found by observing what a man does with it, not what he says about it (6)." Wittgenstein extended this notion and said in his *Philosophical Investigations:* "Meaning of a word is its use." To determine use is to determine its context. Hence the theory is often called the contextual theory of meaning. The technique they use to determine contexts is what is called in modern linguistics the substitution test.

Using the substitution test, in order for us to find the meaning of the term 'cat,' we must find substitutions for such cases as: The _____ caught the mouse; I bought fish for my _____, etc. How many such cases would exhaust all possible occurrences? Suppose we had collected 2,000 such cases, are we to say we have exhausted all possible contexts? How would we know they are all possible contexts? What is more important is that unless we start to *classify* them (2,000 cases), we would not even know how many cases we have collected. Once we start to classify them, however, we are already dealing within the framework of the referential theory of meaning which defines meaning in terms of its relations with others.

It seems clear that some aspects of thesaurus-construction are based on the operational principle: "Meaning is *use*"; therefore, meaning is the sum of all contexts in which a word may occur. (KWIC, for example, is based on this principle.)

Having finished reviewing the two theories of meaning, how do we compare them? Ullmann states:

> The relation between the two methods, or rather between the two phases of the inquiry, is ultimately the same as that between language and speech: the operational theory is concerned with meaning in *speech*, the referential with meaning in *language*. There is absolutely no need to set the two approaches against each other: each handles its own side of the problem, and neither is complete without the other (7).

Which theory of meaning does a thesaurus presuppose? Although we see instances of utilizing the operational theory of meaning in designing a thesaurus, for example, the use of "qualifiers," it seems clear that thesaurus-construction does presuppose primarily the referential theory of meaning.

It is, therefore, in the context of the referential theory of meaning that we defined meaning as: "The reciprocal and reversible relationship between concepts and words." We feel that this is the notion of meaning that a thesaurus presupposes. How then can we organize meaning as we define it? That is precisely the task of theaurus-construction.

Traditionally, efforts to organize meaning took two radically different forms in lexicography. Two radically different forms, however, are explained within the framework of our definition of meaning. See Figure 3. Since the relationship between concepts and words is reciprocal and reversible, we can either start to organize *words* first to get to *concepts* or we can start to organize *concepts* first to get to *words*. The former results in a dictionary; the latter results in a thesaurus (general thesauri such as Roget's.)

If we start with words and look for concepts (designing a dictionary), we are largely in an inductive mode whereas if we start with concepts and look for words (designing a general thesaurus), we are largely in a deductive mode of operation. Neither is perfect without the other. In fact, they are and should be in a complementary relationship.

The shortcomings of using a dictionary are many. For example, the mechanics of alphabetization are tedious and cumbersome. Meaning of a word is given by using more words creating the problem of circularity. For example, upon looking up of the word A, we might be led to the word B, and looking up the word B might lead us to the word C, and looking up the word C might lead us right back to the word A, the meaning of which we are looking for.

The shortcomings of using a general thesaurus are also many. Perhaps the greatest problem is the impression of a "mosaic" it creates. We taxonomize concepts *as if* the entire sphere of human concepts are divided up by different "stones" in a mosaic each of which appears as an idea-word in a thesaurus. The confusion lies in: whose divisions are they? Modern philosophy of science has amply shown that such divisions are man's not nature's.

It seems clear that a thesaurus we use in information storage and retrieval (for convenience, we will call this type a retrieval thesaurus) has elements of both dictionaries and general thesauri. By using "qualifiers" and "scope notes," we are "defining" words in a retrieval

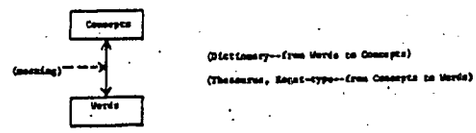


FIG. 3

Reproduced with permission of the copyright owner. Further reproduction prohibited without permission.

thesaurus and by utilizing the faceted display organization, we are indeed "mimicking" general thesauri, such as Roget's.

In any event, when we establish "semantic interrelationships" between terms, it is under these constraints we are operating. That is, if we over-stress "from words to concepts" more than "from concepts to words," we will be merely collecting terms imposing less structure, that is, involving the use of fewer categories. On the other hand, if we over-stress "from concepts to words" more than "from words to concepts," we might be unnecessarily restricting our concepts-sphere by over-categorizing. In the former case, we are displaying a lot of "forms" (as opposed to "substance" to borrow the terms of linguistics) in order not to classify our concept too simply, whereas in the latter case, we are displaying a lesser number of forms, but at the risk of over-simplifying our concept-sphere by classifying concepts. Alphabetic dictionaries are examples of the former while general thesauri such as Roget's example are of the latter.

It is under this kind of tradeoff notion, we must design a thesaurus establishing: ". . . semantically related terms . . ." Just what effect does the way we classify our concept have upon the way we think or perceive the reality? This point, too, will be discussed later when we consider Humboldt.

Theoretically speaking, to summarize, by designing a thesaurus what we are attempting to do is to explicate meaning (the reciprocal and reversible relationship between *concepts* and *words*) in such a way that in the process we may retrieve a subset element (say, a given document or documents) of that meaning. Explicating meaning has to come before the retieval because the retrieval of a "non-relevant" item is worse than none at all.

In designing a thesaurus under our definition of meaning, we employ two fundamentally different approaches: 1) finding words first then looking for concepts; 2) finding concepts first then looking for words. If one approach is over-stressed at the expense of the other, we are, in that case, *not properly* representing meaning in our thesaurus. What we are aiming at is the right mixture of both features taking advantage of the strength of both. (Our next task is to formalize this "theory" which is beyond the scope of this paper.) In short, the retrieval of documents is not the only purpose of using a retrieval thesaurus. As the result of retrieving a document(s), we must acquire some idea as to the meaning it represents, which is more than what is retrieved. Certainly, through the use of a retrieval thesaurus, we retrieve "words" but the retrieval itself should not be the end; what is retrieved should reflect the meaning of which it (the retrieved document) is a part. This is what is meant by "explicating the relationship between concepts and words." This must be the kind of notion Kent had in mind when he stated that the use of a retrieval thesaurus should "enable the user of the information file to predict the relevance of responses to questions. . ."

Let us now briefly discuss the concept of knowledge a retrieval thesaurus seems to presuppose. As we read in the UNESCO *1970-Guidelines*, we seem to be assuming that "a specific domain of knowledge" is always *static* but the vocabulary we use to describe it is *dynamic*. This statement should at best be taken as an hypothesis. We wonder to which specific domain of knowledge we must turn to solve, say, such interdisciplinary problems as "air pollution," "narcotics," etc. Furthermore, what do we mean by "specific?" Is political science, for instance, considered to be a specific domain of knowledge? It is not altogether clear.

Recall carefully what UNESCO says on this point again: ". . . a controlled dynamic vocabulary . . . comprehensively covers a specific domain of knowledge." If they meant, as they seem to in their writing, that "a specific domain of knowledge" is static but vocabularies that cover it are dynamic, one cannot help being reminded of the linguistic Weltanschauung (world-view) theory proposed in the 19th century by Wilhelm von Humboldt. The theory has become, in recent times, known as the Sapir-Whorf hypothesis.

The linguistic Weltanschauung theory is concerned with the way a given language system affects one's perception of the world. Among the first scholars who brought attention to this problem was Wilhelm von Humboldt (1836–39) and Baudouin de Courtenay (1929). Later, in America, Edward Sapir and Benjamin Whorf wrote extensively about this problem. They believed that our categorization of the world (or the way we perceive things) is totally determined by the structure of our native language. "It is language," to quote Leroy on this, "that creates thought: just as numbers help us to calculate so words help us to think (*8*)."

To state it strongly, it is analogous to saying: one cannot think of something if that something is not linguistically expressible. Let us quote Whorf:

> Formulation of ideas is not an independent process strictly rational in the old sense, but is part of a particular grammar and differs, from slightly to greatly, as between different grammars. We dissect nature along lines laid down by our native languages. The categories and types that we isolate from the world of phenomena we do not find there because they stare every observer in the face; on the contrary, the world is presented in a kaleidoscopic flux of impressions which has to be organized by our minds. We cut nature up, organize into concepts, and ascribe significances as we do, largely because we are parties to an agreement to organize it in this way—an agreement that holds throughout our speech community and is codified in the patterns of our language. The agreement is, of course, an implicit and unstated one, but its terms are absolutely obligatory; we cannot talk at all except by subscribing to the organization and classification of data which the agreement decrees (*9*).

In short, they (Humboldt, Sapir, Whorf, and others) thought that language formulates thought instead of the other way around. The question we have on the statement ". . . a controlled dynamic vocabulary . . . com-

Reproduced with permission of the copyright owner. Further reproduction prohibited without permission.

prehensively covers a specific domain of knowledge" is whether or not the vocabulary we choose influences our knowledge, tending to make it *static*. Which one comes first? The static-specific knowledge? or vocabularies? If we were to believe scholars of the linguistic Weltanschauung theory, what we would be saying is that vocabularies formulate our knowledge rather than the other way around. Interpreted differently, the more control we impose on language the more limited our ways of thinking will be. In this case, we are not *discovering* vocabularies that cover a specific domain of knowledge when we are designing a thesaurus; we are *creating* a specific domain of knowledge by controlling vocabularies which in turn allow us to think only in certain ways.

As we will see in the next section on updating, thesaurus-updating is largely done under this principle. Recall Whorf's statement: ". . . we cannot talk at all except by subscribing to the organization and classification of data which the agreement decrees." The way thesaurus-updating is done now, no new terms can enter into a thesaurus unless they ("new terms") are shown to be related to those already in the thesaurus ("old terms"). Another way of stating the situation is: "Unless you are using my terms you are not thinking at all!" We are in fact asserting that knowledge-growth is necessarily incremental in nature. If so, what kind of knowledge are we presupposing when we design a thesaurus? Static? Dynamic? We are not sure. The state-of-the-art of thesaurus-updating seems to subscribe to the Whorf-Sapir hypothesis, which may be interpreted somewhat strongly as: "If you choose certain words and categorize them in certain ways, you are allowing users of a system to think only in certain ways."

It might be noted, however, that Whorf-Sapir's hypothesis is believed by many scholars to be *largely untenable* although some interesting studies on color terms and kinship terms seem to support the hypothesis.

In any event, what kind of concept of knowledge does a retrieval thesaurus assume or presuppose? We can see two features: 1) the *static* knowledge as shown in the UNESCO *1970-Guidelines;* 2) the "incremental" nature of knowledge in that no "new" knowledge is possible which is *not* based on "old" knowledge. (This point is proven by the way a retrieval thesaurus is updated today. For example, for most of the retrieval thesauri, no candidate terms are accepted for inclusion unless they are shown to be related to those terms already in the thesaurus—"you are not thinking unless you think in my terms.") Are these two features of knowledge representative of the prototype of knowledge that we strive to achieve or gain? Maybe the all-elusive Sapir-Whorf hypothesis should be tested in the area of retrieval thesaurus-construction, application, and maintenance. For example, some of the questions may be: just how much *control* do we need to foster "the healthiest" thinking on the user's part? Just how much classifying of vocabularies in a thesaurus will best allow the growth of knowledge on the user's part? and so on.

• The State-of-the-Art of Updating Thesauri

As we pointed out, because of our poor (or lack of) understanding of the theoretical foundations underlying rules and conventions of construction, we have no definitive statements as to how a thesaurus should be systematically updated.

The purpose of this section is to report on *how a thesaurus is actually updated*. To this end, a survey was conducted (*10*). Two sets of questionnaires were designed to find out: 1) How is a thesaurus actually updated? 2) If updated at all, who does it?—by input people (indexers) or by output people (question negotiators). The first set of questionnaires was sent to input people and the second set of questionnaires was sent to output people. The working hypothesis was that if a thesaurus was not updated by both input and output people, then the communication process between authors and readers is essentially a one-way (or linear as opposed to cybernetic) communication. See Figure 5.

Survey Design and Findings

The sample size: 120 operating systems in the "hard" sciences and behavioral sciences were selected from *The Encyclopedia of Information Systems and Services* (ed. Kruzas, 1971).

The selection procedures: systems listed in the *Encyclopedia* under the headings "behavioral sciences," "sociology," and "political science," 51 of them altogether, were identified. If a system was listed more than once but under different headings, the system was counted only once. Of 51 systems identified in behavioral sciences, 18 systems claimed that they provide indexing services or perform their own indexing and 33 systems claimed that they provide searching and retrieval services without performing their own indexing.

In addition, of the systems listed in the *Encyclopedia* under the headings "science and technology" and "engineering," 69 operating systems were identified. Here, too, duplication was avoided.

Of 69 systems operating in "hard" sciences, 25 systems which provide indexing services or perform their own indexing and 44 systems which provide searching and retrieval services without performing their own indexing were identified. (See Figure 4)

In sum, one set of questionnaires was sent to 43 indexers (18 indexers of behavioral sciences plus 25

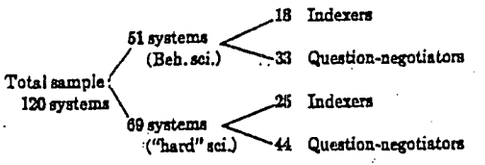


Fig. 4

Reproduced with permission of the copyright owner. Further reproduction prohibited without permission.

indexers of "hard" sciences) and another set of questionnaires was sent to 77 question-negotiators.

Of 120 systems which received the questionnaire, 61 systems responded to the questionnaire (about 50% return). Of 61 systems which responded, 44 systems (72.1%) reported "No" to the question: "Do you update or revise, or participate in updating or revising, the thesaurus you employ? That is, do you recommend terms for thesaurus revision if they are not found in the thesaurus, but feel they should be included?" This means that only 29.9% of the systems that responded perform some kind of updating.

Probably the most important finding is the fact that out of 17 systems (29%) which claimed updating, 12 of them (about 70%) were from input people and only 5 were from output people. This could mean that the communication between authors and readers that we (system designers) aim to make most effective could be linear ("one-way") rather than cybernetic in nature, supporting our original working hypothesis. (See Figure 5.) The working hypothesis was: if a thesaurus is updated by only one party, that is, only by input people (II in Fig. 5) or output people (IV in Fig. 5) then the communication between authors and readers is linear rather than cybernetic. One of the important concepts of a cybernetic process is the notion of feedback which means, for our purpose, that a thesaurus must be updated by both input and output people if the communication process between authors and readers is to be cybernetic rather than linear. We are here, of course, assuming that in updating a thesaurus the input people (indexers) reflect authors' point of view while the output people (question-negotiators) reflect readers' point of view. However, if an author (or authors) and a reader (or readers) are the same person, we can achieve a cybernetic process of communication in spite of the fact that a thesaurus is updated by a single party (either input or output people). It seems that the so-called "scientific thesauri" are built and maintained on this notion. (For example, a chemist writes for other chemists.)

• Suggestions for Possible Research-Techniques to be Utilized for Systematic Updating of Thesauri.

What are the reasons for updating a thesaurus? We have already discussed indirectly one reason from the communications point of view—to make the communication process cybernetic. But are there any linguistic and semantic reasons for updating thesauri?

Perhaps the following quotation from Leroy explains aptly why we should update thesauri:

> Phonological and grammatical elements are in most cases closely organized and relatively stable in a given state of the language, whereas the vocabulary is essentially unstable: it is constantly enriched with new words while others fall into disuse or are given new meanings (*11*).

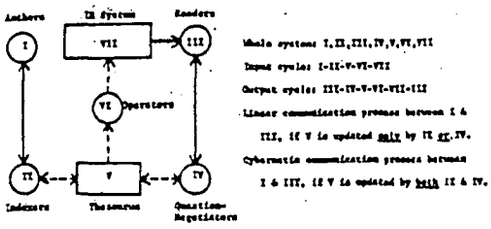


FIG. 5

Vocabularies are coming and going. Our tasks are to tell "when" and "how." But then, what is meant by "update?" It seems clear from the above quotation that by "update" we mean three things: 1) the addition of new words with new meanings; 2) the addition of new meanings for old words; 3) the deletion of words.

Since we discussed why one should update a thesaurus and itemized what the term "update" means, let us now consider some possible techniques which, if applied, will enable us to perform the systematic updating. We can see at least two different methods or techniques useful for updating.

They are: 1) the content analysis technique based on an etymological approach; and 2) the semantic field theory and its techniques.

Thorndike in his article "Semantic changes" (*12*) has discussed how one can measure quantitatively the rate of the addition of new words with new meanings, the rate of addition of new meanings for old words and the rate of dropping-out because of disuse. This study was based on dates as given in the Oxford English Dictionary for the period of about 600 years, i.e., dates of entries and exits of some 10,000 words were carefully counted and plotted to detect trends of semantic changes. He literally counted the "ages" of words. We believe that the same kind of technique may be utilized for our purpose to learn *what* kinds of words need to be updated and when.

Let us now consider *how* to update. Maybe we can learn something from the technique used in semantic field theory. The technique used in semantic field theory, which is the second of the two methods we think to be useful for updating thesauri, is based on the association principle of words. There are different versions of field theories but for our purpose Charles Bally's field concept seems to be most useful. He defines a field as: "an intricate network of associations (of words), some based on similarity, others on contiguity, some arising between senses, others between names, others again between both (*13*)." Bally envisioned the associative field as a group of terms which surrounds a particular term or terms like a halo. In this halo, no one term is given meaning independent of other words in the same halo. That is, the meaning of a word within a field is given only in relation with others in the same field. However, the association is such that one term

Reproduced with permission of the copyright owner. Further reproduction prohibited without permission.

may occur in more than one field. In other words, fields do overlap each other. What criteria must one use to identify the elements of a "halo" objectively enough to identify the central associations? Words for such fields as military rank, color terms, and kinship terms were quite convincingly demonstrated, however, terms for abstract ideas present different problems.

Bally and his associates used what may be called the "interview" technique, i.e. asking subjects to say synonyms, antonyms, and homonyms given a "stimulus" word.

Whether we accept the semantic field theory or not, there is still a lesson to be gained, i.e., we are building a thesaurus the same way Bally conceptualized his "halo." In a thesaurus, as in Bally's halo, the meaning of a word is given in relation to others. In other words, the meaning of a word is given in the context of associative nets within a halo. For example, "Broader Term," when used in a thesaurus, is "broader" only in relation to its hyponyms which are its "Narrower Term(s)." Also in a thesaurus, as in Bally's notion of halo, a term may occur more than once, forming different associative nets. For example, the vertical control on associative nets is achieved by "Broader Term," "Narrower Term," etc. while the horizontal control is achieved by "Related Term."

What can we do with this technique? Once we have established the rate of adding and deleting words using Thorndike's method, we will then have some knowledge as to *what* to update and *when*. Once we know this, we can then pick terms from a thesaurus to establish "new" associations as Bally did for his terms by asking subjects for most obvious synonyms and antonyms as well as homonyms. Likewise, once we know the trend for the occurrence of new words and their rate of changes, we will at least know where to look for "new" terms. If we do this applying Thorndike's method and the field theory technique, we will have indeed combined the quantitative-etymological technique with the structural approach to the study of meaning. This has never been done and we think it is worth an attempt.

• Summary and Conclusion

It has been argued that the present-day retrieval thesaurus-construction and maintenance rules and conventions are not theoretically based. Consequently, systematic rules and conventions for updating or maintaining a thesaurus are few, if any; thus with the present state-of-the-art one continues to construct "new" thesauri whenever "old" ones cease to function effectively.

Thesaurus-updating was viewed from a communications point of view. The working hypothesis was that the communication process between authors and readers is linear ("one-way" only) if a thesaurus adopted in a system is updated by either indexers or question-negotiators alone. Stated differently, the communication process is cybernetic ("both-way" allowing feedback) if a thesaurus is updated by both indexers and question-negotiators. A survey was conducted to test the hypothesis. Findings show that most of the retrieval thesauri adopted today in operating information-storage and retrieval systems are not updated at all. Where updating is done, it is carried out mostly by indexers alone which indicates that the use of a retrieval thesaurus allows a linear process of communication between authors and readers, allowing no feedback.

No attempt was made to develop a theory of retrieval thesaurus-construction and maintenance. It was, however, strongly argued that a theory, if developed, should account for the concepts of meaning and knowledge. In other words, the practice of thesaurus-construction and maintenance ought to be based on at least two theories: a theory of meaning and a theory of knowledge. With this challenge in mind, the concepts of meaning and knowledge that present-day thesauri seem to presuppose was discussed. To reiterate, the challenge that a theory of thesaurus-making and updating must meet is to account for the *conversion of what is read or written into knowledge and the conversion of knowledge into what is read or written*. This is the reason we have defined meaning as the reciprocal and reversible relationship between concepts and words.

References

1. UNESCO. *Guidelines for the Establishment and Development of Monolingual Scientific and Technical Thesauri for Information Retrieval.* Paris. 1970.
2. COSATI. *Guidelines for the development of information retrieval thesauri.* Washington: Govt. Print. Office. 1967.
3. KENT, ALLEN. *Information Analysis and Retrieval.* New York: Wiley. 1971. p. 230.
4. KENT, ALLEN. "Controlling the Output of Mechanized Retrieval Systems: an Empirical Test of the NASA Thesaurus," Unpublished report to Chief, Dissemination Branch (Code UTD). Technology Utilization Division. NASA Sept. 1969.
5. ULLMANN, STEPHEN. *Semantics: An Introduction to the Science of Meaning.* Oxford. 1970. p. 57.
6. CHASE, STUART. *The Tyranny of Words.* New York: Harcourt, Brace and Company. 1938. pp. 11–12.
7. ULLMANN. *op. cit.* p. 67.
8. LEROY, MAURICE. *Main Trends in Modern Linguistics.* Berkeley, Calif.: Univ. of Calif. Press. 1967. p. 26.
9. WHORF, BENJAMIN. (reprinted from) *The Technology Review.* M.I.T. 1939. p. 231.
10. KIM, CHAI. "How are Thesauri Actually Updated?—A survey." Unpublished Internal paper. IDIS. University of Pittsburgh. 1972.
11. LEROY. *op. cit.* p. 133.
12. THORNDIKE, EDWARD. "Semantic Changes," *The American Journal of Psychology.* v. 60 (1947) pp. 588–97.
13. ÖHMAN, SUZANNE. "Theories of the 'Linguistic Field,'" *Word.* v. 8 (1953) p. 126.

Reproduced with permission of the copyright owner. Further reproduction prohibited without permission.

Bibliography

ANGELL, R. S. *Two Papers on Thesaurus Construction.* Copenhagen: Danish Center for Documentation. 1968. 27 p.

ATHERTON, P. A. "Is Compatibility of Authority Files Practicable?" (In: Newman, S. M. ed. *Information Systems Compatibility.* Spartan. 1965. pp. 69–80.)

BARHYDT, GORDON. *Information Retrieval Thesaurus of Education Terms.* Case Western Reserve University. 1968. 133 p.

BLAGDEN, J. G. "Thesaurus Compilation Methods: A Literature Review." Aslib (Association of Special Libraries and Information Bureaux) *Proceedings.* v. 20 (Aug. 1968). pp. 345–59.

DAVIS, C. H. *Approach to Automated Vocabulary Control in Indexes of Organic Compounds.* Thesis Ph.D. Indiana University. 1969. 125 p.

DYM, E. "A new Approach to the Development of a Technical Thesaurus." American Documentation Institute. Annual Meeting. *Proceedings.* 1967. pp. 126–31.

ERIC. *Rules for Thesaurus Preparation.* ERIC. Sept. 1969.

Engineers Joint Council. *Thesaurus of Engineering and Scientific Terms:* a list of engineering and related scientific terms and their relationships for use as a vocabulary reference in indexing and retrieving technical information. New York: Engineers Joint Council. 1967. Appendix 1 "Thesaurus Rules and Conventions."

FARRADANE, J. E. L. and others. "Problems in Analysis and Terminology for Information Retrieval." *Journal of Documentation.* v. 21 (Dec. 1965) pp. 287–90.

FUJIKAWA, M. "Characteristics and use of Thesauri." *Library Science.* (Mita Society of Library Science). no. 4. 1966. pp. 29–42.

GASTER, K. (comp.) "Thesaurus Construction and Use, a Selective Bibliography based on material in the Aslib Library in July 1967." Aslib (Association of Special Libraries and Information Bureaux) *Proceedings.* v. 19 Sept. 1967). pp. 310–17.

GOLDWYN, A. J. "The Vocabulary Control Problem." *Educational Research.* Supplement. 1967.

GREER, ROYAL. "User Vocabulary in Thesaurus Development." *Perceptual and Motor Skills.* v. 21 (1965). pp. 827–37.

HERSEY, D. F. and W. HAMMOND. "Computer Usage in the Development of a Water Resources Thesaurus." *American Documentation.* v. 18 (Oct. 1967). pp. 209–15.

HOHNECKER, WALTER and MARK NEWMARK. "Automated maintenance of a highly structured Thesaurus at Engineering Index." American Documentation Institute. Annual Meeting. *Proceedings.* 1967. pp. 132–36.

HINES, T. C. and J. L. HARRIS. "Columbia University School of Library Service Systems for Thesaurus Development and Maintenance." *Information Storage and Retrieval.* v. 7 (June 1971). pp. 39–50.

JOYCE, T. and R. M. NEEDHAM. "The Thesaurus Approach to Information Retrieval." American Documentation. v. 9 (July 1958). pp. 192–97.

KEEN, E. MICHAEL. *Thesaurus, Phrase and Hierarchical Dictionaries.* (In: Cornell University. Department of Computer Science. Information Storage and Retrieval. Scientific Report no. ISR–13 to the National Science Foundation. Reports on Evaluation Procedures and Results 1965–67. Gerard Salton, Project Director.) Ithaca, N.Y. Dec. 1967. 59 p.

KRIŠ, CHAI. "Can Men find the 'Meaning' of Meaning?" *ETC: A Review of General Semantics.* v. 29 (Sept. 1972). pp. 251–55.

KOCHEN, M. and R. TOGLIOZZO. "Study of Cross-Referencing." *Journal of Documentation.* v. 24 (Sept. 1968). pp. 173–91.

MANDERSLOOT, W. G. and others. "Thesaurus Control: The Selection, Grouping, and Cross Referencing of Terms for Inclusion in a Coordinate Index Word List." *Journal for American Society for Information Science.* v. 21 (Jan. 1970). pp. 49–57.

MARGOSI, A. "EURATOM Thesaurus and UDC combined for the Subject Organization of a small Information Service." *Journal of Documentation.* v. 25 (Sept. 1969). pp. 197–213.

MARTINEZ, S. J. and D. P. HELANDER. "Development and Maintenance of Specialized, Controlled Vocabulary Thesaurus." American Society for Information Science. Conference. *Proceedings.* 1968. pp. 279–83.

MULVIHILL, J. G. and E. H. BARNNER. "Faceted Organization of Thesaurus Vocabulary." American Documentation Institute. Annual Meeting. *Proceedings.* 1966. pp. 175–83.

NEVILLE, H. H. "Feasibility Study of a Scheme for Reconciling Thesauri Covering a Common Subject." *Journal of Documentation.* v. 26 (Dec. 1970). pp. 313–36.

PRICE, SAMUEL T. "The Development of a Thesaurus of Descriptors for an Information Retrieval System in Special Education." Unpublished Ph.D. Thesis. University of Pittsburgh. 1969.

ROLLING, L. N. "Compilation of Thesauri for Use in Computer Systems." *Information Storage and Retrieval.* v. 6 (Oct. 1970). pp. 341–50.

ROSTRON, R. M. "Construction of a Thesaurus." Aslib (Association of Special Libraries and Information Bureaux) *Proceedings.* v. 20 (Mar. 1968). pp. 181–87.

SOERGEL, D. "Some Remarks on Information Languages, their Analysis and Comparison, with Comment by P. A. Atherton and R. A. Fairthorne. *Information Storage and Retrieval.* v. 3 (Dec. 1967). pp. 219–96.

SPARCK, JONES, K. "Automatic Thesaurus Construction and the Relation of a Thesaurus to Indexing Terms." Aslib (Association of Special Libraries and Information Bureaux) *Proceedings.* v. 22 (May 1970). pp. 226–28.

VICKERY, B. C. "Thesaurus—A New Word in Documentation." *Journal of Documentation.* v. 16 (Dec. 1960). pp. 181–89.

WALL, E. *Information Retrieval Thesauri.* New York: Engineers Joint Council. 1962. 17 p.

WALL, E. "Vocabulary Building and Control Techniques." *American Documentation.* v. 20 (Apr. 1969). pp. 111–4.

Reproduced with permission of the copyright owner. Further reproduction prohibited without permission.

The Application of a Minicomputer to Thesaurus Construction
Edward John Kazlauskas; Thomas D Holt
Journal of the American Society for Information Science (pre-1986); Sep 1980; 31, 5; ABI/INFORM Global
pg. 363

The Application of a Minicomputer to Thesaurus Construction

Edward John Kazlauskas
School of Library Science, University of Southern California, Los Angeles, CA 90007

Thomas D. Holt
VORT Corporation, P.O. Box 11132, Palo Alto, CA 94306

The use of a minicomputer in various phases of creating the thesaurus for the National Information Center for Special Education Materials (NICSEM) database is described. The minicomputer is used to collect, edit, and correct candidate thesaurus terms. The use of the minicomputer eases the process of grouping terms into files of similar concepts and facilitates the generation of products useful in vocabulary review and in term structuring. Syndetic relations, indicated by assigning coded identification numbers, are altered easily in the design phase to reflect restructuring requirements. Because thesaurus terms are already in machine-readable form, it is simple to prepare print programs to provide permuted, alphabetic, hierarchical, and chart formatted term displays. Overall, the use of the minicomputer facilities initial thesaurus entry development by reducing clerical effort, editorial staff decisions, and overall processing times.

Introduction

In October 1977 a contract was awarded to the University of Southern California by the Bureau of Education for the Handicapped (BEH) to continue the development of a system which would allow access to and retrieval of information on instructional materials appropriate for special education. A requirement for the development of this National Information Center for Special Education Materials (NICSEM) was the creation of a search vocabulary that would describe learner performance in terms of skill-oriented objectives. The initial task of selecting and structuring this vocabulary was assigned to a subcontractor, VORT Corporation of Palo Alto, California, which worked closely with NICSEM information science and special education staff.

The search term vocabulary for NICSEM was to be controlled by a thesaurus. Unfortunately, no universally accepted terminology or structure existed in the field of special education which could be used to build this thesaurus. There were many developmental taxonomies and behavioral systems that had been designed to identify and help meet the needs of special education students, but no single system or set of terms had gained uniform acceptance. However, these existing taxonomies and behavioral systems, as well as other sources such as standardized tests, could be used for thesaurus term generation. Likewise, a parallel empirical strategy could be utilized to identify terms from actual instructional materials.

Gathering terms from such multiple sources to generate a universe of terms appropriate to the subject area is a recognized step in thesaurus construction and was used for the NICSEM system. This process is complicated, and made particularly so since the domain of special education overlaps with many other disciplines such as medicine, general education, and psychology. Typically, the manual generation of a controlled vocabulary from diverse sources involves voluminous paper processing and clerical time for tasks such as typing, editing, and sorting. These tasks appeared ideally suited for computerization. Additional advantages of computerization were evident in other areas of thesaurus work such as the generation of term reciprocals.

It should be noted that the use of computers in thesaurus construction is not a new application. Artandi [1] refers to the use of the computer in preparing permuted types of indexes, such as KWIC, and in manipulating natural language text to produce a vocabulary. The manipulation of natural language text, or automatic language processing, is described in detail in the literature by others such as Salton [2]. Another use of computers in vocabulary work is that performed by PRECIS wherein the computer processes an input string provided by the indexer in order to generate index entries and references [3]. Recently, minicomputer-operated information retrieval systems have been developed which include, for example, CAIRS [4] and MORPHS [5].

However, the NICSEM project did not require these applications but rather something akin to what Hines [6] describes as computer-assisted index and thesaurus entry development. Several reports on this aspect of computerized

Received January 17, 1980; accepted March 21, 1980.

© 1980 by John Wiley & Sons, Inc.

0002-8231/80/0031-0363$01.00

Reproduced with permission of the copyright owner. Further reproduction prohibited without permission.

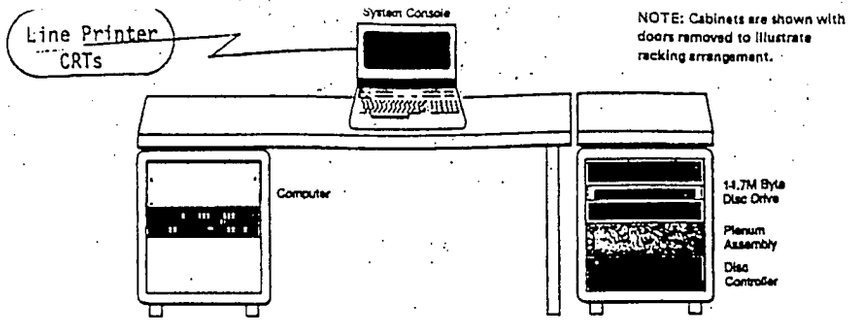


FIG. 1. HP 1000 model 30 computer system.

thesaurus work appear in the literature. Schultz and Shepherd [7] describe the use of the computer to compile a machine-readable thesaurus for the 1960 Federation of American Societies for Experimental Biology Meeting. Hersey and Hammond [8] provide a description of the use of the computer in the development of a *water resource* thesaurus, while Martinez [9] describes the computer generation of a controlled vocabulary for *Petroleum Abstracts*. In addition, the introductions to several published thesauri indicate the use of the computer in thesaurus construction. For example, Schultz describes the computer techniques used to prepare the *RSA Research Information System Thesaurus* [10] and the *Thesaurus of Information Science Terminology* [11]. However, these references tend to describe early computer processing techniques, normally batch processing on medium-sized computers using punched card input. Advances in computer processing, particularly with regard to the online/interactive mode, have increased the type and number of "housekeeping" functions that could be computerized. The use of the computer in an online mode would assist in generating a universe of terms which could be machine manipulated and hierarchically structured. Additionally, reciprocal relationships could be generated automatically to provide visual/printed displays of segments of the term file and thus greatly assist in vocabulary review and in statistical and historical analysis [12].

Computer Characteristics

Computer processing of the type just mentioned required several capabilities: online term input; online editing, sorting, and grouping; high-speed text analysis for building the term database; and different printing formats for the separate thesaurus indexes.

VORT Corporation had recently acquired a Hewlett-Packard (HP) 1000 minicomputer (Fig. 1) to be used primarily for internal accounting and product development applications, including text processing and manipulation similar to that necessary for the new thesaurus development process.

As a minicomputer, the HP 1000 had a high-speed processor (595 nsec), an effective, multitasking operating system (RTE), and a moderate amount of online disc storage (15 mbytes). FORTRAN had been selected as the programming language to provide fast execution and avoid the overhead (time/space) of a database management system. The data entry and processing capabilities of the HP 1000 made it an excellent choice for creating the thesaurus database. The decision to use the HP 1000 was made by VORT after an initial analysis of data flow suggested that an expedient was necessary to meet the short time frame of the initial thesaurus term generation project.

The speed of FORTRAN along with direct access files proved to be a major asset in processing/adding new terms and in merging the terms into alternate indexes within an online environment. The powerful HP 1000 editor permitted updating of the FORTRAN programs concurrent with the emerging thesaurus design requirements. The actual thesaurus development programs were adapted by the subcontractor from existing software and ran in less than 10K words of user memory with a common driver/overlay structure.

Development Process

After an initial review of other thesauri, a development strategy was formulated for construction of the NICSEM Thesaurus [13]. The important considerations included the use of field "experts" to review and edit the terms, the comprehensive study of special education literature and sources of terms, and the formatting of the final product in a manner most functional for the user—the teacher. The overall activities associated with this strategy divided into five basic segments (Fig. 2); these included (1) term collection, (2) term editing/synthesis, (3) term grouping, (4) term structuring, and (5) index printing. The capabilities of the online thesaurus programs developed for the HP 1000 were utilized throughout segments 2-5.

Term Editing/Synthesis

The term collection process yielded an initial 16,000 terms. These terms were collected based solely on their possible reference to special education and not necessarily their applicability to skill-oriented objectives. Identified terms were recorded on either worksheets or highlighted

Reproduced with permission of the copyright owner. Further reproduction prohibited without permission.

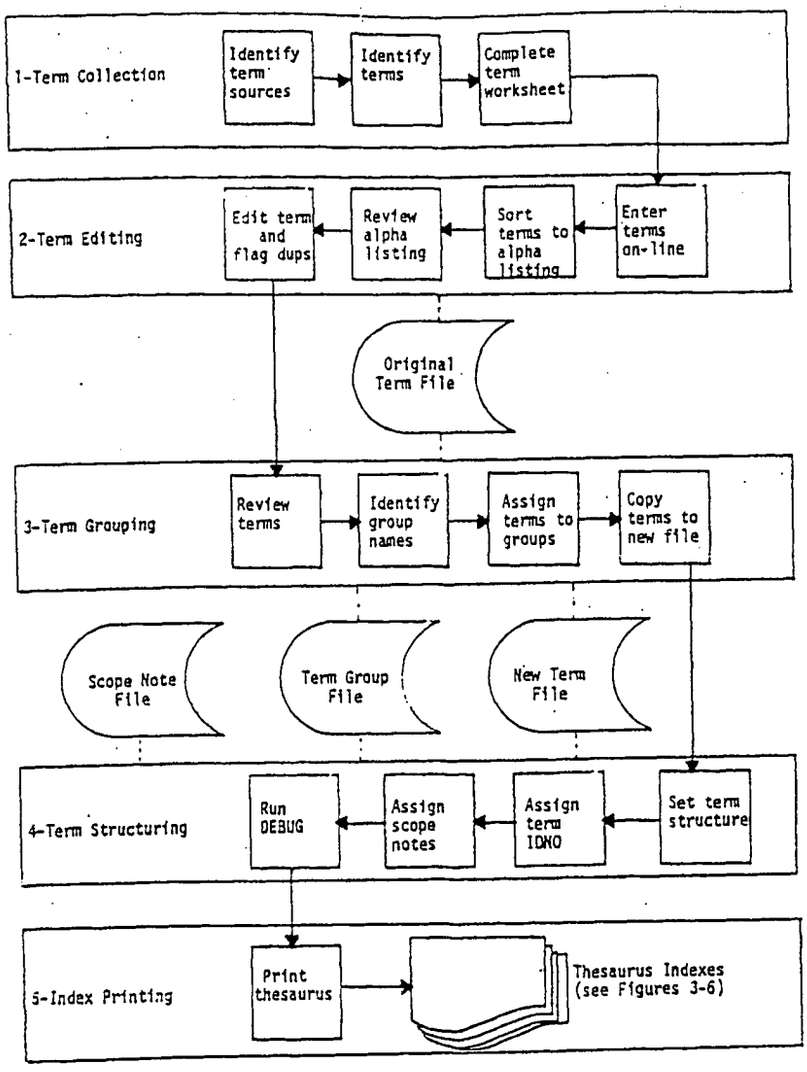


FIG. 2. NICSEM Thesaurus development process

directly in the source documents. The programs for developing the thesaurus ran from one main program (driver). All the input programs were tutorial in nature. They provided the entry operator with complete instructions at each step of the entry process to assure accuracy and efficiency. The input programs appeared as a series of branching options which could be selected by the entry operator according to a variety of functions. To enter a term, the operator would sign on the CRT and select the option titled "new term." The system would ask the operator for the necessary term data including text, source number, and scope note if available. The operator would enter the data via the CRT in response to each question. The CRTs ran at 9600 baud and provided the operators with a high-speed entry process which averaged more than 250 terms/hr.

During the entry process, terms were not edited nor were duplicates avoided or deleted. The entry operator merely entered the data as they appeared on the source document. Entry was not verified. If operators made typographical errors, they could exit the current option, and select the "change term" option to directly update/correct any term previously entered. On a periodic basis, typically every one or two days, all the terms in the term file were sorted and printed in alphabetic sequence. This alphabetic listing was used by the editorial staff to identify duplicate terms, correct errors, and verify term form. Duplicate terms were not physically deleted from the term file but were flagged. A specific field within the term's logical record was set to a certain value if the term was a duplicate or determined during the review of the listing to be inappropriate for special education. Subsequent alphabetic listings did not include flagged terms. Since terms were flagged and retained in the original term file, they could be reviewed and reconsidered throughout the development process. This was an important feature because it allowed the editorial staff greater flexibility in assigning and manipulating terms.

Reproduced with permission of the copyright owner. Further reproduction prohibited without permission.

```
Checking Accounts
Savings Accounts
Accreditation
Achievement Tests
Career Acquisition
Acting
Action Verbs
Reflective Or Active Listening
Active Passive Voice
Enrichment Activities
Leisure Activities
Playground Activities
Recreational Activities
Self Actualization
Visual Acuity
Adaptive Behavior Scales
Adaptive Communication Devices
Adaptive Eating Devices
Adaptive Travel
Adding Machine Use
Addition
Addition Facts
Addition Regrouping
Column Addition
```

FIG. 3. NICSEM permuted index—by VORT Corp.

```
BB          Exceptionalities and Disorders
BB-03         Arithmetic Disorders
BB-06         Behavior Disorders
BB-09         Birth Problems
BB-0903         Birth Defects
BB-0906         Premature Birth
BB-12         Communication Disorders
BB-1203         Agnosia
BB-1206         Anomia
BB-1209         Aphasia
BB-1209.03        Receptive Aphasia
BB-1209.06        Expressive Aphasia
BB-1212         Auditory Processing Disorders
BB-1215         Dysnomia
BB-1218         Dysphasia
BB-1218.03        Expressive Dysphasia
BB-1221         Expressive Language Disorders
BB-1224         Language Disorders
BB-1227         Memory Disorders
BB-1230         Receptive Language Disorders
BB-15       Deaf Blind Individuals
BB-18       Disadvantaged Individuals
BB-21       Emotional Disturbances
BB-2103       Autism
BB-2106       Neurosis
BB-2109       Psychosis
BB-2112       Schizophrenia
BB-2115       Serious Emotional Disturbances
BB-24       Giftedness
```

FIG. 5. NICSEM hierarchical index—by VORT Corp.

The term-editing process was completely independent of the term-entry process and permitted the editorial staff to concentrate on the term form while the entry operators could, on a parallel basis, concentrate on data entry. A separate list of terms which were constantly appearing as duplicates was provided to the entry operators so that they could do some minor preediting to minimize the duplicate entry of high-occurrence terms.

During the term-editing process, the staff began to identify the similar concepts/clusters associated with the various terms. This was not part of the formal term grouping effort, but provided an initial analysis of the terms by group. Names of potential groups were identified and discussed by the editorial staff in preparation for segment 3 of the process, term grouping. As terms were edited and flagged, approximately 13,000 terms of the original 16,000 were left unflagged and available as "candidate terms."

```
XH-03       Abandoned Children
HJ-0924     Abbreviations
XR-03       Ability Tests
XL-03       Abnormal Psychology
NJ-0303     Abolition
FX-0609.03  Abortions
              SN: Sex education
ZZ-1050     Abstract Reasoning
              Use: Reasoning
YJ-0303     Academic Degrees
YF-02       Academic Education
ZZ-2415     Academic Giftedness
              Use: Giftedness
YJ-0306     Academic Rank
              SN: Professional
ZZ-2365     Acalculia
              Use: Arithmetic Disorders
HF-1512.15  Accents
FJ-24       Accessibility Skills
              SN: Ability to recognize and remove or overcome an
                  obstacle which may prevent ones entrance to or
                  use of a building or piece of equipment
ZZ-0510     Accident Prevention
              Use: Safety
ZZ-0415     Accident Procedures
```

FIG. 4. NICSEM alpha index—by VORT Corp.

Term Grouping

After several weeks of term collection and editing, the editorial staff used the most recent alphabetic listing along with previous notes on possible term groups to identify term group names. The thesaurus design required that the term groups relate to specific skills for special education students. However, during the study of the updated alphabetic listing, it became apparent that many of the terms were not skill related and yet appropriate for special education. Some of the terms related to program administration, academic studies, or type of special education category. Term groups were, therefore, sorted into skill-oriented and nonskill categories. After proposed names for the various term groups were selected, an online group file was created. This file was to contain the name of each group and the term numbers in the original term file relating to that group. The text of each term remained in the original file, and only the direct access pointer (address) was stored in the group file. The input option "term assignment" was used to assign terms by number to the group file. Another option, "change groups," could be used to add or delete term numbers to a specific group or to change the name of the group. All candidate terms, i.e., those not flagged, were assigned initially to 150 groups. Through continuous editing and indexing actual instructional materials, term groups were revised until approximately 100 groups remained containing the numbers of only 5000 of the original 16,000 terms. This left 11,000 terms in the original file with flags.

The remaining 5000 candidate terms were copied to a newly created term file for faster scanning, editing, and sorting. The review process was continued and 2300 more terms were flagged, leaving 2700 usable terms.

Reproduced with permission of the copyright owner. Further reproduction prohibited without permission.

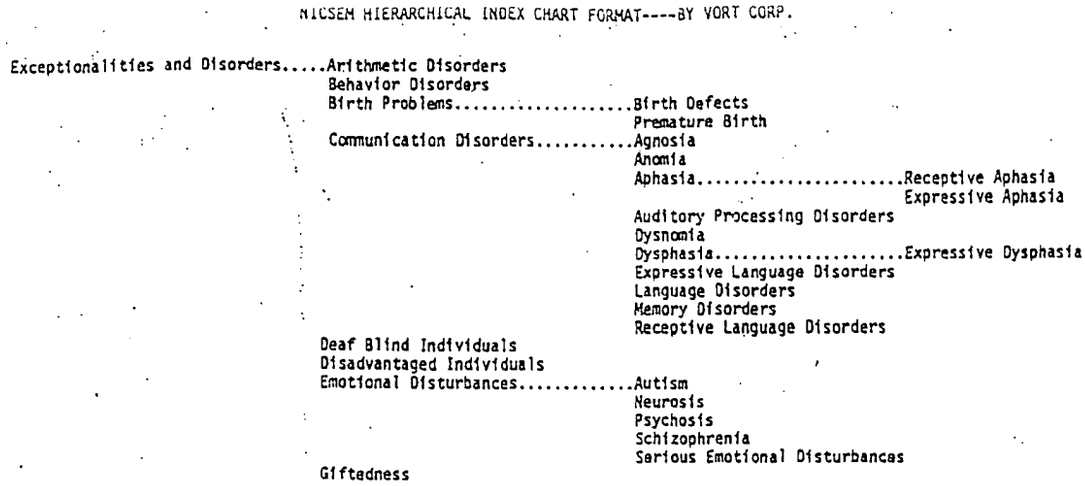


FIG. 6. NICSEM hierarchical index chart format—by VORT Corp.

Term Structuring

The usable terms were reviewed by group, and a five-level alphanumeric identification number (IDNO) was assigned online to each term. Based on input from experts and sampling of materials, those terms that were synonyms were assigned IDNOs indicating that they were unacceptable and should be used as reciprocals to acceptable terms. Acceptable terms were assigned IDNOs which placed them in their appropriate position within the hierarchy, and in like manner pointed to unacceptable reciprocals. Sampling of domain materials was continued resulting in terms being added and deleted from groups. If required, term IDNOs could be altered to accommodate these changes. In this manner, group structure was synthesized and new group listings printed of each group. Terms were also printed by level across groups to help refine the overall structure. For example, the IDNO permitted displaying all level 1 terms, or alternately, displaying all level 2 terms as one group to help compare term form and position within the hierarchy.

Index Printing

After refinement of the overall group structure and assignment of IDNOs, terms were printed in alphabetic sequence using a special format that permitted the writing of scope notes directly on the output document. Parenthetical qualifiers were also edited using this format. The online file maintenance programs permitted rapid entry of these scope notes into a separate scope note file related to the term file by direct access pointers. A separate scope note file was used to save file space since not all terms were to have scope notes. A "debug" program was run against the term file to verify reciprocal relationships between all acceptable terms (Use For-UF) and their reciprocal unacceptable terms (Use Reference—UR). These steps yielded term and scope note files containing 2060 UF and 560 UR terms.

Three separate programs were run against the term and scope note files to print the final subcontract products: a permuted index (Fig. 3), an alphabetical index (Fig. 4), and a hierarchical index (Fig. 5). A fourth print program was also prepared to print portions of the hierarchical index in a chart format (Fig. 6). The samples shown in Figures 3–6 reflect interim products developed by VORT for NICSEM and do not represent the final or official published version of the NICSEM thesaurus. It should be noted that NICSEM is now adding additional terms and term relationships to the thesaurus. These are being generated from the actual indexing process.

Conclusion

In the creation of the NICSEM thesaurus, the term entry, editing, and structuring processes were greatly enhanced by using the minicomputer. Specifically, terms were collected and entered into the system quickly. The online, direct access files permitted editorial staff to interact directly with term data. The speed and capabilities of the computer system reduced editorial staff decision and processing time by an estimated 30%. The use of the HP 1000 replaced the need for a large number of clerical staff to prepare and update term listings on a manual basis.

The use of the minicomputer in thesaurus construction allowed great staff flexibility and a standard continuity to the term synthesis. The experience in generating the NICSEM thesaurus indicates that the minicomputer can provide an economic and replicable process [14] for thesaurus construction not available through manual or large, batch-oriented systems.

References

1. Artandi, S. A. "Machine Indexing: Linguistic and Semiotic Implications." *Journal of The American Society for Information Science.* 27(4):235-329; 1976.
2. Salton, G. *The SMART Retrieval System; Experiments in Automatic Document Processing.* Englewood Cliffs, NJ: Prentice-Hall; 1971.

Reproduced with permission of the copyright owner. Further reproduction prohibited without permission.

3. Austin, D.; Digger, J. A. "PRECIS: The Preserved Context Index System." *Library Resources and Technical Services.* 21(1): 13-30; 1977.
4. Bell, C. L. M.; Jones, K. P. "A Minicomputer Retrieval System with Automatic Root Finding and Rolling Facilities." *Program.* 10(1):14-27; 1976.
5. Bell, C. L. M.; Jones, K. P. "Towards Everyday Language Information Retrieval Systems via Minicomputers." *Journal of The American Society for Information Science.* 30(6):334-339; 1979.
6. Hines, T. "Computer-Assisted Index and Thesaurus Entry Development." *Proceedings of The American Society for Information Science.* 10:94-95; 1973.
7. Schultz, C. K.; Shepherd, C. A. "The 1960 Federation Meeting: Scheduling a Meeting and Preparing an Index by Computer." *Federation of American Societies for Experimental Biology Proceedings.* 19:682-699; 1960.
8. Hersey, D. F.; Hammond, W. "Computer Usage in The Development of a Water Resource Thesaurus." *American Documentation.* 18(4):209-215; 1967.
9. Martinez, S. J.; Brown, L. P.; Helander, D. P.; McLeod, H. "Computer Processing of Thesaurus Data." *Proceedings The American Society for Information Science.* 6:269-27 1969.
10. Schultz, C. K. Compiler and Ed. *RSA Research Informati System Thesaurus.* Washington, D.C.: U.S. Department Health Education and Welfare, Office of Human Developme Rehabilitation Services Administration, Office of Research a Demonstrations; 1975.
11. Schultz, C. K. *Thesaurus of Information Science Terminolog rev, ed.* Metuchen, NJ: Scarecrow Press; 1978.
12. Lancaster, F. W. *Vocabulary Control for Information Retriev* Washington, D.C.: Information Resources Press; 1972: 90-9
13. *NICSEM Special Education Thesaurus.* Los Angeles, CA: Un versity of Southern California, The National Informatio Center for Educational Media; 1979.
14. This same thesaurus construction process was used to genera the *Chicano Thesaurus for Indexing Chicano Materials.* San Barbara, CA: University of California; 1979.

Reproduced with permission of the copyright owner. Further reproduction prohibited without permission.

علم.ر.٢٩٦

Information Processing & Management Vol. 27, Nos. 2/3, pp. 165–175, 1991
Printed in Great Britain.

0306-4573/91 $3.00 + .00
Copyright © 1991 Pergamon Press plc

SPECIFICATIONS FOR THESAURUS SOFTWARE

JESSICA L. MILSTEAD*
The JELEM Company, Brookfield, CT 06804, U.S.A.

(*Received* 23 *April* 1990; *accepted in final form* 11 *July* 1990)

Abstract—Specifications are presented for software designed to support manual development and maintenance of information retrieval thesauri. The specifications are intended to support both evaluation of existing packages for acquisition and design of custom software. Specialized requirements of thesaurus support are emphasized over more general database management requirements. Requirements for integration with larger systems and for the user interface are summarized. Relationships specified by the ANSI standard should be supported, and the relationships should be validated to avoid introduction of conflicting relationships. Flexibility in availability of note and date fields is important, and where required, term classifications should be available. Maintenance procedures should include provision for feedback on the impact of changes, as well as appropriate levels of approval for changes. A wide variety of online displays and printed reports are required, including but not limited to alphabetical, hierarchical, and rotated. In addition, it is desirable to be able to search the thesaurus database with such capabilities as Boolean logic and proximity operators.

INTRODUCTION

Considering the large number of thesauri that have been and continue to be produced, the general lack of software designed specifically for the requirements of thesaurus databases is surprising. Several new packages or modules have become available in the past year or two, however, and it seems likely that more will be arriving soon. The capabilities of these packages vary, but none fully meets the needs of even fairly standard thesauri, let alone specialized situations.

The primary purpose of this paper is to provide guidance in designing and evaluating thesaurus software. The goals are to suggest useful features to designers, and to aid prospective purchasers in defining their needs and setting priorities. No attempt is made to provide general specifications for database design. Eddison (1985) provides a general checklist which is extremely useful as an aid in defining requirements for a database management system. This paper concentrates on features that are more specialized to thesauri.

Little has been published on specifications for development or selection of thesaurus software. Wall's (1989) report was prepared in the context of a particular firm's needs. The draft revised National Information Standards Organization (NISO) (1990) standard for thesauri contains general recommendations for the capabilities of thesaurus software. All of its recommendations have been incorporated into this paper, as have the recommendations on thesaurus structure found in the present American National Standard (1980). Milstead (1990) has described the capabilities of existing software, both independent packages and modules of larger systems. Reference may be made to that paper for information on the degree to which individual systems meet the specifications discussed here.

The specifications presented in this paper are derived from the sources referenced, from examination of numerous thesauri, and from work with several thesaurus packages. No attempt is made to present a prescriptive list to be used in selecting software. Recommendations are made for some criteria; for others, the alternatives are presented, to aid in consideration of the needs of a particular situation. An attempt is made, particularly for more important criteria, to indicate the degree to which available software meets a specific need.

Address for all communications: P.O. Box 5063, Brookfield, CT 06804, USA

Supplied by the British Library "The world's knowledge"

- Initial input of data should be carried out quickly, with a minimum of waiting for machine processing.
- Any batch processing requires sophisticated conflict prevention routines. For example, if one side of a cross reference is input, the system must generate the reciprocal. If, however, one side is deleted, the system must not later detect the fact that there is an unreciprocated reference (the side that was not deleted), and restore the deleted reference. Instead, it must either record the fact that the reciprocal is to be deleted, or delete it immediately.
- Batch processing routines must carry out all validation and conflict prevention that are also available for online input.
- Because most errors detected in batch must be reported rather than corrected automatically, and because in any large run one error may affect another, repetitive batch runs are usually necessary to produce a clean file. The more errors that can be caught in online preprocessing, the better.

Security

The security requirements for thesaurus databases are not unique, but the different needs of developers and system administrators, indexers, and searchers must all be provided for.

Limits

Many systems limit the length of terms and/or notes. There may also be limits on the number of occurrences of a multi-occurrence field, or on the total number of relationships assigned to a given term. When such limits exist, it is important to verify that they are generous enough to meet the need. A limit on descriptor length should not be less than 60 characters; 100 is better. If proper names are to be included in the thesaurus, an even higher term length will be required. There should be no limit on the length of a note; these can easily run to hundreds of characters.

Customization

The degree of customization required varies with the complexity of the application. It is important to define system requirements well enough to determine the extent to which a proposed system can be customized to meet them. Some kinds of needs that should be part of the initial implementation are:

- Specialized relationship types
- Note fields
- Dates
- Term approval procedures
- Reports and displays
- Screen layout
- Menu design.

Needs will change over time, as well; therefore it is important to determine the extent to which thesaurus maintenance staff can make changes without going back to systems staff. Some examples of later changes that will certainly be needed are:

- New and specialized reports
- Changed report formats
- Procedures for term approval
- Authorizing and de-authorizing users
- New note fields.

Sorting

Sorting for either online display or reports should be case-insensitive, with upper- and lower-case characters not differentiated. Although it is true that an acronym (AID) and a

Supplied by the British Library - "The world's knowledge"

- Generation of such aids as permuted, hierarchical, or alphabetic lists based on a requested term
- Searching with Boolean logic, proximity search, etc.

Either of these methods is likely to be adequate. Searching of the usual database type is more likely to be available with software that is a module of a larger DBMS. The following functions are crucial:

- Ease of locating the record for a term without knowing a code number
- The ability to locate all terms containing a given word
- The ability to move either from the record for a term to that of one of its related terms, or to a term record from a listing without keying the new term (e.g., by highlighting or keying a menu number).

Error messages

These should describe the type of error as clearly as possible, and should be in plain English. They also must stay on the screen long enough to be read comfortably—preferably until a keystroke calls for the next action from the system. There are few things more frustrating to users than to turn away from the screen while waiting for processing to complete, only to discover that an error message has appeared and gone while the screen was not being watched.

INTEGRATION WITH LARGER SYSTEM

Thesauri are not used in isolation, and an important criterion in selecting software is the ease with which the thesaurus can be integrated with the database(s) for which it is the controlled vocabulary. Software modules are usually seamlessly integrated with their parent DBMS, but are unlikely to be directly usable with any other system. The few standalone packages vary in the ease with which they can be integrated with other software.

Indexing use

In indexing, such functions as the following should be supported:

- Validation of indexer input, to prevent indexing with unauthorized or misspelled terms
- Addition of candidate terms to the thesaurus during the indexing process
- Automatic switching (preferably with notification) from nonpostable to postable terms for both indexing and search
- Review of a term's relationships, to consider other terms for use in indexing or search.

Validation of indexer input can go a step further, with the indexer selecting the desired indexing term (by number, clicking on it, etc.) without keying it. If this procedure is made compulsory, then wrong terms cannot be input. Great care in design would be required, however, to assure that this step did not slow down indexing; finding the term and inserting it into the field should take no more time than keying it in the first place.

Thesaurus maintenance and database update

Full integration makes it possible for a change in a thesaurus term (as long as it is a one-for-one change) to cause a global change to the database, so that all documents indexed by the old version will be changed to the new version of the term.

Search use

Online access to the thesaurus can also be of great value to searchers, permitting them to examine thesaurus terms and their relationships, and to include terms in their searches without rekeying them. The thesaurus can be used as an entry vocabulary, switching the

Supplied by the British Library - "The world's knowledge"

Reciprocation of relationships

Reciprocation of relationships must be automatic. Of the relationships described above, the user should input one side, relying on the system to generate the other side. While the reciprocal tags for standard relationships should be predefined, for user-defined relationships these tags must be user-specifiable. A reciprocal is necessary for every relationship for control and maintenance purposes, even if not all are to be printed in end-user versions of a thesaurus. For instance, the reciprocals (X and XX) of SEE and SEE ALSO relationships frequently are not shown in subject heading lists.

A further form of reciprocation, called up-posting, is used in some thesauri. Terms are posted as NTs to every term above them in the hierarchy, not just the next level upward. The purpose is to show the full hierarchy of a term at every level. However, up-posting destroys the logical arrangement. In a thesaurus of any size it will also produce unmanageably long lists of BTs and NTs. Hierarchical displays are a better means of showing full hierarchies. Nonetheless, if up-posting is needed in a particular situation, the software must provide for it.

Validation

As noted above, validation may take place online, in batch, or in a combination of the two modes. Crucial to the quality of a thesaurus is that the validation procedure exist, be accurate, and be executed as required, whether automatically or upon demand. There are two aspects of validation:

- Validation of terms
- Validation of relationships

Term validation

When a relationship is added to the thesaurus, the system must verify that the term referred to is indeed in the system, exactly as shown. In order to avoid introduction of typographical errors and synonymous variants, a new main term should not be generated automatically on the basis of being shown in a relationship to another term.

When an attempted relationship does not match an existing term, there are two basic choices for the system response:

- Rejection of the relationship, requiring that the term be added as a separate step if it is indeed valid (the only option in batch mode)
- A response showing that the term is not now in the thesaurus, and querying whether it should be added.

The first of these options is more fail-safe, because it prevents a possibly erroneous or conflicting term from being added casually without further checking. However, it also makes adding terms and relationships more cumbersome, particularly during initial development when many new terms and relationships are being added. The second option is not as fail-safe, but is more user-friendly. Terms cannot be added accidentally, but it is possible, while working on one term, to input another one into the system for later refinement.

The best compromise is probably a variant of the second option. The maintainer is notified that the term is not now in the thesaurus, and is allowed to add it to a candidate term file. The candidate terms are then processed before being admitted to the thesaurus itself. This allows for easy recording of terms as they arise, while holding them aside for full analysis before generating relationships and including them in the thesaurus proper.

Relationship validation

The first check in relationship validation is the automatic combination of duplicates. Particularly if a large number of terms and relationships are added in batch mode during early development, there will be instances where both sides of a relationship are input; routine generation of reciprocals would result in two "copies" of the relationship under each

Supplied by the British Library - "The world's knowledge" www.bl.uk

- Authority note: recording sources searched for the term
- Deletion/modification note: recording for all users the period during which an obsolete term was used for indexing, or the date when a new term was introduced, and the term used for the topic during other periods.

Dates

As a minimum, the date when a term was added to the thesaurus must be recorded. It is desirable to add other dates, including dates of proposal and approval, as well as of the most recent modification, or even of all modifications since the term was added. Because terms may be dropped from active use in indexing without being removed from documents in the database, a date of deletion is valuable as well.

THESAURUS MAINTENANCE

Feedback

All transactions with the thesaurus should receive feedback to confirm that the change was made and that it is correct. In addition, it should not be too easy to delete a term record. An erroneous change in a relationship may be detected later, but when a term is gone there is no record of it to show the gap. As a minimum, the system must prompt for confirmation that the deletion is intended. Prompting may go even further, for instance, to remind the user that deleting a term deletes all its relationships as well. If deleting a term will produce an orphan (a term with no relationships), this fact should also be signaled.

Approvals

Procedures for approval of both new terms and new relationships need to be provided. The person responsible for approval may not be the one doing the keyboarding. Even if a single person carries out all these operations, it is wise to determine the actual impact of an addition or change before finally integrating it into the database.

Orphans

An orphan is a term with no relationships. Although it is desirable that every term be part of a hierarchy (i.e., that it have at least one BT or one NT), this is not always possible. At the very least, however, an RT should be present. Software packages vary in their treatment of orphans: some do not permit them to be included in the database; others accept them without comment. Orphans should be accepted; particularly in the development stages, it is often desirable to add a term, returning later to include its relationships. Otherwise, it can be cumbersome to begin adding terms in a new subject area.

Global changes

There will be times when it is necessary to change the same feature of many terms, for example, a change in spelling policy, or a shift in categories. A global change facility makes this possible. Complete reporting of every transaction is essential for global changes, so that unintended changes can be corrected.

REPORTS AND DISPLAYS

A wide variety of screen displays and printed reports are needed; in general, any arrangement available on-screen should be printable, and vice versa. Five general kinds of displays should be considered: alphabetical, hierarchical, rotated, microthesaurus, and maintenance.

Alphabetical

Frequently the alphabetical listing is the most complete; it usually shows all the fields associated with a term, plus all its relationships. Both postable and nonpostable terms will be included.

Supplied by the British Library - "The world's knowledge" www.bl.uk

- Validation of terms in relationships
- Support of all ANSI/NISO-standard relationships
- Availability of additional free-text fields
- Confirmation of changes
- A wide variety of reports and displays.

Acknowledgement—Carole Schaub provided information that was of aid in preparation of this paper.

REFERENCES

American National Standards Institute (1980). *Guidelines for Thesaurus Structure, Construction and Use*. New York: American National Standards Institute. (Z39.19–1980).

Eddison, B. (1985). Database Design: A Checklist for Defining Your Needs. *Database*, *8*(3), 79–81.

Milstead, J.L. (1990). *Thesaurus Software Packages*. In Proceedings of the 53rd Annual Meeting of the American Society for Information Science, Toronto, Canada, November 1–5, 1990. Medford, NJ: Learned Information, Inc. (In press).

National Information Standards Organization (1990). *Proposed American National Standard Guidelines for Thesaurus Structure, Construction and Use*. Gaithersburg, MD: National Information Standards Organization. (ANSI/NISO Z39.19–199x).

Wall, E. (1989). *Vocabulary Management: A Technical Report*. 5th edition. Rockville, MD: Aspen Systems Corp.

PERGAMON

Information Processing and Management 36 (2000) 361–378

INFORMATION PROCESSING & MANAGEMENT

www.elsevier.com/locate/infoproman

Query expansion using heterogeneous thesauri

Rila Mandala*, Takenobu Tokunaga, Hozumi Tanaka

Department of Computer Science, Tokyo Institute of Technology, 2-12-1 Ookayama Meguro, Tokyo, 152-8552, Japan

Received 23 February 1999; accepted 27 October 1999

Abstract

This paper proposes a method to improve the performance of information retrieval systems by expanding queries using heterogeneous thesauri. The expansion terms are taken from a hand-crafted thesaurus, co-occurrence-based automatically constructed thesaurus, and predicate-argument-based automatically constructed thesaurus. To avoid the effects of wrong expansion terms, a weighting method is devised such that the weight of expansion terms depend not only on the weight of all terms in query, but also the weight of those terms in each thesaurus. Experiments show that using heterogeneous thesauri with an appropriate weighting method results in better retrieval performance than using only one type of thesaurus. © 2000 Elsevier Science Ltd. All rights reserved.

Keywords: Query expansion; Thesauri; Text retrieval

1. Introduction

One major problem in information retrieval is the difficulty in describing user information needs in terms of a query so that a system can accurately distinguish between relevant and irrelevant documents for the query (Salton & McGill, 1983). In fact, the user's original query statement will usually consist of just a few terms related to the subject of interest.

Query expansion is a technique utilized within information retrieval to remedy this problem (Ekmekcioglu, 1992; Fox, 1980). A query is expanded by adding other terms that are closely related to the original query terms. Expansion terms can be selected by referring to thesauri (Crouch, 1990; Crouch & Yang, 1992; Jing & Croft, 1994; Kristensen, 1993; Paice, 1991) or by

* Corresponding author. Tel.: +81-3-5734-2831; fax: +81-3-5734-2915.
E-mail address: rila@cs.titech.ac.jp (R. Mandala).

0306-4573/00/$ - see front matter © 2000 Elsevier Science Ltd. All rights reserved.
PII: S0306-4573(99)00068-0

consulting users through the relevance feedback technique (Salton & Buckley, 1990; Buckley & Salton, 1994). Past research has verified the effectiveness of relevance feedback, but it puts a burden on users to a certain extent. Furthermore, if a user is not familiar with the vocabulary of a document collection, it is difficult to obtain good expansion terms, unless the system can suggest terms to the user.

Currently, many information retrieval researchers use pseudo relevance feedback, which is relevance feedback without user intervention (Buckley & Salton, 1995; Xu and Croft, 1996). Firstly, several documents are retrieved as a result of the initial retrieval. Assuming that the top-n retrieved documents are relevant, the system uses the terms contained in those documents as expansion terms and retrieves again. This technique works well when the system has a fairly good performance and the top-n retrieved documents are really relevant. If this assumption does not hold, however, the performance of the system can degrade as the expansion terms are taken from irrelevant documents.

Query expansion using thesauri is still worth investigating. Briefly, there are two types of thesauri, that is, hand-crafted thesauri (Fox, 1980) and automatically constructed thesauri (Chen, Schatz, Yim & Fye, 1995; Crouch, 1990; Crouch & Yang, 1992). WordNet (Miller, Beckwidth, Fellbaum, Gross & Miller, 1990) is an example of a hand-crafted thesaurus which is available in machine readable form. Many researchers have tried to use relations defined in WordNet for query expansion. Unfortunately, the results have not been as good as expected. Voorhees (1994) used WordNet as a tool for query expansion. She conducted experiments using the TREC collection in which all terms in the queries were expanded using a combination of synonyms, hypernyms, and hyponyms. She set the weight of the words contained in the original query to 1, and used a combination of 0.1, 0.3, 0.5, 1, and 2 for the expansion terms. She then used the SMART information retrieval system (Salton, 1971) to retrieve the documents. Through her experiments, Voorhees succeeded in improving the system performance on only short queries with little or no significant improvement for long queries. She further tried to use WordNet as a tool for word sense disambiguation (Voorhees, 1993) and applied it to text retrieval, but the performance of retrieval was degraded.

Smeaton, Kelledy and O'Donnel (1996) tried to expand the queries of the TREC-4 collection with various strategies of weighting expansion terms, along with manual and automatic word sense disambiguation techniques. Unfortunately, all strategies degraded the retrieval performance.

Unlike hand-crafted thesauri, corpus-based thesauri are constructed automatically from corpora without human intervention. There are two different methods of extracting thesaural relationships from corpora, that is, co-occurrence statistics (Chen et al., 1995; Crouch, 1990; Qiu & Frei, 1993; Schutze & Pedersen, 1994) and grammatical relations (Grefenstette, 1992, 1994; Hindle, 1990; Jing & Croft, 1994; Ruge, 1992).

Qiu and Frei (1993) used an automatically constructed thesaurus and improved retrieval effectiveness by about 20% using small test collections. Schutze and Pedersen (1994, 1997) also built a co-occurrence-based thesaurus and applied it to two information retrieval applications. Using a scaled-down TREC collection, he slightly improved the retrieval performance.

Peat and Willett (1991) have provided theoretical evidence of the limitation of term co-occurrence data for query expansion in information retrieval. Consequently, some researchers have tried to build thesauri using more linguistically motivated methods. Ruge (1992) built a

linguistically-based thesaurus, but she did not apply it to information retrieval. Grefenstette (1992) built a thesaurus using syntactic context and performed experiments using several small test collections. His method improved the performance for some small collections, but failed to improve the performance using other collections (Grefenstette, 1994). Jing and Croft (1994) also found an improvement through query expansion by using a grammatically-based automatically constructed thesaurus.

Despite the success of using corpus-based thesauri (Qiu & Frei, 1993; Jing & Croft, 1994; Schutze & Pedersen, 1997), the improvement is very limited when compared to the relevance-feedback method. This paper tries to analyze why query expansion using a thesaurus shows only limited performance, and based on this analysis we propose a new method to improve the performance of query expansion by combining heterogeneous thesauri for query expansion.

2. Analysis of thesaurus characteristics

As stated earlier, query expansion using WordNet has not always succeeded in improving the performance of information retrieval systems. One reason for this would be that WordNet is a general-purpose thesaurus. It has been compiled to be used in a wide variety of domains, and therefore lacks domain-specific thesaural relationships. We conducted query expansion experiments using WordNet and found limitations, which are summarized as follows:

- Two interrelated words may have different parts of speech. This is the case with "stochastic" (adjective) and "statistics" (noun). Since words in WordNet are grouped on the basis of part of speech, it is not possible to find a relationship between terms with different parts of speech.
- Most domain-specific relationships between two words are not found in WordNet as mentioned above.
- Some kinds of words, such as proper names, are not included in WordNet.

On the other hand, corpus-based thesauri are able to incorporate domain specific information, since they can be automatically constructed from a set of documents relating to a certain domain. Some researchers have succeeded in improving the performance of information retrieval systems with query expansion using corpus-based thesauri (Qiu & Frei, 1993; Jing & Croft, 1994; Schutze & Pedersen, 1997).

The construction of thesauri generally takes the following form (Charniak, 1993):

1. Extract word co-occurrences.
2. Define similarities (distances) between words on the basis of co-occurrences.
3. Cluster words on the basis of similarities.

Depending on the context of "co-occurrence", different types of thesauri can be constructed. In past attempts, document co-occurrence relations and predicate-argument relations have often been used as the co-occurrence context.

A drawback of using a document co-occurrence relation is that it is difficult to capture relations between certain words that share the same meaning. For example, "tumor" and "tumour" denote the same concept, but would never appear in the same document, at least not

frequently enough so as to be recognized by a statistical method. This is because one tends not to use a variety of words denoting the same meaning in a single document. This kind of relationship can be found in a general purpose thesaurus such as WordNet, and can also be found in a thesaurus constructed using predicate-argument relations, since we can expect that "tumor" and "tumour" appear as an argument of the same predicate.

Since each type of thesaurus has different advantages and disadvantages, combining them provides a valuable tool for query expansion. In this paper, for the purpose of query expansion, we combine three types of thesauri: a hand-crafted general purpose thesaurus (WordNet), an automatically constructed thesaurus based on a document co-occurrence relation (co-occurrence-based thesaurus) and an automatically constructed thesaurus based on a predicate-argument relation (predicate-argument-based thesaurus).

3. Method

In this section, we first describe the construction method for each type of thesaurus utilized in this research, and then describe a term weighting method using similarity measures based on these thesauri.

3.1. WordNet

WordNet is a hand-crafted thesaurus developed by a Princeton University group led by George Miller (Miller et al., 1990). In WordNet, words are organized into taxonomies where each node is a set of synonyms (a "synset") representing a single sense. There are four different taxonomies based on different parts of speech and also there are many relationships defined among them. In our experiments we use only a noun taxonomy with hyponymy/hypernymy relations (or an *is–a* relation). Fig. 1 is a fragment of the WordNet taxonomy, where *is–a* relations are represented by the ⇒ symbol.

The similarity between words a and b can be defined as the shortest path from each sense of a to each sense of b, as below (Resnik, 1995):

$$\mathrm{sim}_{ab} = \max\left[-\log\left(\frac{N_p}{2D}\right)\right]$$

```
tumor, tumour, neoplasm
      => growth
           => illness, malady, sickness
                => ill health, unhealthiness, health problem
                     => physiological state
                          => condition, status
                               => state
```

Fig. 1. A fragment of WordNet taxonomy.

where N_p is the number of nodes in path p from a to b and D is the maximum depth of the taxonomy. We normalize the similarity value so that its value ranges from 0 to 1.

3.2. Co-occurrence based thesaurus

This method is based on the assumption that a pair of words that occur frequently together in the same document are related to the same subject. Therefore, word co-occurrence information can be used to identify semantic relationships between words. Suppose two words a and b occur in df_a and df_b documents, respectively, and co-occur in df_c documents, then the similarity between a and b can be calculated using a similarity coefficient such as the Dice coefficient (Peat & Willett, 1991)

$$\frac{2 \times df_c}{df_a + df_b}.$$

3.3. Predicate-argument-based thesaurus

In this method term, relations are gathered on the basis of linguistic relations and not document co-occurrence statistics (Hindle, 1990). Words appearing in a similar grammatical context are assumed to be similar, and, therefore, classified into the same class. As grammatical relations, we use subject–verb, verb–object and adjective–noun relations in order to build word classes.

First, all the documents are parsed using the Apple Pie Parser, which is a probabilistic chart parser developed by Sekine (Sekine & Grishman, 1995). Then the following syntactic structures are extracted.

- Subject–verb.
- Verb–object.
- Adjective–noun.

Each noun has a set of co-occurring verbs and adjectives, and for each such relationship, a Dice coefficient value is calculated.

- $C_{\text{sbj}}(v_i, n_j) = [2 \times f_{\text{sbj}}(v_i, n_j)/f(v_i) + f_{\text{sbj}}(n_j)]$, where $f_{\text{sbj}}(v_i, n_j)$ is the frequency of noun n_j occurring as the subject of verb v_i, $f_{\text{sbj}}(n_j)$ is the frequency of noun n_j occurring as the subject of any verb, and $f(v_i)$ is the frequency of verb v_i.
- $C_{\text{obj}}(v_i, n_j) = [2 \times f_{\text{obj}}(v_i, n_j)/f(v_i) + f_{\text{obj}}(n_j)]$, where $f_{\text{obj}}(v_i, n_j)$ is the frequency of noun n_j occurring as the object of verb v_i, $f_{\text{obj}}(n_j)$ is the frequency of noun n_j occurring as the object of any verb, and $f(v_i)$ is the frequency of verb v_i.
- $C_{\text{adj}}(a_i, n_j) = [2 \times f_{\text{adj}}(a_i, n_j)/f(a_i) + f_{\text{adj}}(n_j)]$, where $f_{\text{adj}}(a_i, n_j)$ is the frequency of noun n_j being modified by adjective a_i, $f_{\text{adj}}(n_j)$ is the frequency of noun n_j being modified by any adjective, and $f(a_i)$ is the frequency of adjective a_i.

We define the similarity of two nouns with respect to a predicate, as the minimum of the following Dice coefficients:

$$\mathrm{sim}_{\mathrm{sbj}}(v_i, n_j, n_k) = \min\{C_{\mathrm{sbj}}(v_i, n_j), C_{\mathrm{sbj}}(v_i, n_k)\}$$

$$\mathrm{sim}_{\mathrm{obj}}(v_i, n_j, n_k) = \min\{C_{\mathrm{obj}}(v_i, n_j), C_{\mathrm{obj}}(v_i, n_k)\}$$

$$\mathrm{sim}_{\mathrm{adj}}(a_i, n_j, n_k) = \min\{C_{\mathrm{adj}}(a_i, n_j), C_{\mathrm{adj}}(a_i, n_k)\}.$$

Finally, overall similarity between two nouns is defined as the average of all the similarities between those two nouns for all predicate-argument structures.

3.4. Expansion term weighting method

A query q is represented in terms of a vector $\mathbf{q} = (w_1, w_2, \ldots, w_n)$, where n is the total number of terms in the collection, w_i is the weight of term t_i when t_i is contained in query q. If t_i does not appear in the query, w_i becomes 0.

The similarity between a query q and term t_j can be defined as follows (Qiu & Frei, 1993):

$$\mathrm{sim}_{qt}(q, t_j) = \sum_{t_i \in q} w_i \cdot \mathrm{sim}(t_i, t_j)$$

where $\mathrm{sim}(t_i, t_j)$ is defined as the average of the similarities of the three types of thesauri mentioned in the previous section.

With respect to a query q, all the terms in a collection can be ranked according to their sim_{qt}. Terms t_j with high sim_{qt} in rank are used as expansion terms.

The weight $w_{\mathrm{ex}}(q, t_j)$ of an expansion term t_j with respect to a query q is defined as the following function of $\mathrm{sim}_{qt}(q, t_j)$:

$$w_{\mathrm{ex}}(q, t_j) = \frac{\mathrm{sim}_{qt}(q, t_j)}{\sum_{t_i \in q} w_i}.$$

The weight of an expansion term depends both on all terms appearing in a query and on the similarity between the terms, and ranges from 0 to 1. This weight can be interpreted mathematically as the weighted mean of similarities between term t_j and all terms in the query. The weight of the original query terms are weighting factors of those similarities.

A query q is expanded by adding the vector

$$\mathbf{q}_{\mathbf{e}} = (w_1', w_2', \ldots, w_n')$$

to the original query vector, where w_j' is equal to $w_{\mathrm{ex}}(q, t_j)$ if t_j is in the top r ranked terms, otherwise w_j' becomes 0.

4. Experiments

In order to evaluate the effectiveness of the proposed method, we conducted experiments using the TREC-7 information retrieval test collection (Voorhees & Harman, 1999). The

TREC-7 test collection consists of 50 topics (queries) and 528,155 documents from several sources: the Financial Times (FT), Federal Register (FR94), Foreign Broadcast Information Service (FBIS) and the LA Times. Each topic consists of three sections, the "Title", "Description" and "Narrative". Table 1 shows statistics of the document collection, Table 2 shows statistics of the topics, and Fig. 2 shows an example of the topic.

For the query, we use, respectively, the title only, the description only, and all sections of the topic. Note that in the TREC-7 collection, the description section contains all the terms in the title section.

As a baseline we used the SMART version 11.0 (Salton, 1971) system without query expansion. SMART is an information retrieval engine based on the vector space model in which term weights are calculated based on term frequency, inverse document frequency, and document length normalization. The weighting method for document collection is as follows:

$$\frac{(\log(tf_{ik}) + 1.0)}{\sqrt{\sum_{j=1}^{n}[\log(tf_{ij} + 1.0)]^2}}$$

and the weighting method for the initial query is as follows:

$$\frac{(\log(tf_{ik}) + 1.0) * \log(N/n_k)}{\sqrt{\sum_{j=1}^{n}[\log(tf_{ij} + 1.0) * \log(N/n_j)]^2}}$$

where tf_{ik} is the occurrence frequency of term t_k in query q_i (for query terms weighting) or in document d_i (for document terms weighting), N is the total number of documents in the collection, and n_k is the number of documents to which term t_k is assigned. This term weighting method is called the *lnc.ltc* weighting method.

The results are shown in Table 3. This table shows the average of non-interpolated precision for each case, expansion using only WordNet, expansion using only the predicate-argument based thesaurus, expansion using only the co-occurrence based thesaurus, and expansion using the combination of all thesauri types. For each method we give the percentage of improvement over the baseline method in parentheses. We demonstrate that the performance using the

Table 1
TREC-7 document statistics

Source	Size (MB)	No. of docs	words/doc. (median)	words/doc. (mean)
Disk 4				
FT	564	210,158	316	412.7
FR94	395	55,630	588	644.7
Disk 5				
FBIS	470	130,471	322	543.6
LA Times	475	131,896	351	526.5

Table 2
TREC-7 topic length statistics (words)

Topic section	Min.	Max.	Mean
Title	1	3	2.5
Description	5	34	14.3
Narrative	14	92	40.8
All	31	114	57.6

combined thesauri for query expansion is better than SMART version 11.0 using the *lnc.ltc* term weighting method without expansion and than expansion using just one type of thesaurus.

5. Analysis of results

5.1. Contribution of each thesaurus

We investigated to what extent each thesaurus contributes in providing expansion terms. Table 4 summarizes the percentage of expansion terms in expanded queries that are suggested by one thesaurus only, two of the thesauri, and all thesauri. We can see that each thesaurus contributes almost the same in providing expansion terms. Table 5 shows example expansion terms and their source thesauri, for the query: *how often were the peace talks in Ireland delayed or disrupted as a result of acts of violence*?

5.2. Different coefficient measures

We also investigated the effect of different coefficient measures for constructing thesauri.

Title
commercial cyanide uses

Description
What are the industrial or commercial uses of cyanide or its derivatives?

Narrative
A document is relevant if it names or describes a process that uses cyanide commercially or mentions that cyanide-rich waste comes from a particular industry.

Fig. 2. Topic example.

Table 3
A comparison of the average non-interpolated precision for baseline, single, pairwise, and combined thesauri

Topic type	Base	Expanded with						
		WordNet only	Pred-arg only	Co-occur only	WordNet + pred-arg	WordNet + co-occur	Pred-arg + co-occur	Combined method
Title	0.1175	0.1276 (+8.6%)	0.1473 (+25.4%)	0.1603 (+36.4%)	0.1611 (+37.1%)	0.1698 (+44.5%)	0.1859 (+58.2%)	0.2213 (+88.3%)
Description	0.1428	0.1509 (+5.7%)	0.1626 (+17.1%)	0.1846 (+29.3%)	0.1832 (+28.3%)	0.1973 (+38.2%)	0.2315 (+62.1%)	0.2590 (+81.4%)
All	0.1976	0.2010 (+1.7%)	0.2203 (+11.5%)	0.2333 (+18.1%)	0.2276 (+15.2%)	0.2423 (+22.6%)	0.2565 (+29.8%)	0.2573 (+30.2%)

- Mutual information measure (MI)

  The mutual information measure compares the probability of the co-occurrence of words a and b with the probabilities of the occurrence of a and b independently (Hindle, 1990). The definition of the mutual information measures is as follows:

$$\mathrm{MI}(a, b) = \log \frac{P(a, b)}{P(a) \cdot P(b)}$$

  The probabilities of $P(a)$ and $P(b)$ are estimated by counting the number of occurrences of a and b in document collection and normalizing by the total number of words in the document collection. The value of $P(a, b)$ is estimated by counting the number of times that word a co-occurs with b and is also normalized.
- Tanimoto coefficient

  Suppose two words a and b occur in df_a and df_b documents, respectively, and co-occur in df_c documents, then the similarity of a and b can be calculated using the Tanimoto coefficient (Peat & Willett, 1991) as follows:

$$\frac{df_c}{df_a + df_b - df_c}.$$

Fig. 3 shows a recall-precision graph using the description section of the topics, and Table 6 shows the average of non-interpolated precision for different topic sections using different similarity measures. Although the mutual information measure gives better performance than the other coefficient measure methods, the difference is not significant.

5.3. Number of expansion terms

Fig. 4 shows the effect of the number of expansion terms added to the original query, on retrieval performance. The retrieval performance is measured by the average non-interpolated precision over all relevant documents. The improvement in expansion increases when adding up to 20 terms, and reaches a plateau, then the improvement begins to decrease when more than 50 terms are added. The expanded queries, however, still perform better than the original queries. The results shown in Fig. 4 also indicate that expanding a query with 30 to 40 top ranked terms seems to be the safest method with respect to the collection targeted in this evaluation.

Table 4
Percentage of expansion terms in the expanded query suggested by given source thesauri

Source thesauri	Title (%)	Description (%)	All (%)
WordNet	6	7	5
Co-occurrence	12	9	20
Pred-arg	10	8	14
WordNet and co-occurrence	14	13	9
WordNet and pred-arg	13	13	7
Co-occurrence and pred-arg	17	28	32
All	28	22	13

5.4. Size of corpora for thesaurus construction

We now describe an experiment to see the effect on retrieval performance of the size of corpora which thesauri are constructed from. In other words, we would like to see whether the combined thesaurus needs to be updated when new documents are added to the collection. We first built thesauri using only TREC-7 disk 4, and performed retrieval experiments using disks 4 and 5. Figs. 5–7 show the improvement in retrieval performance with expanded queries using these thesauri over the original queries using the title only, the description only, and all parts of the topic, respectively. For comparison, we also show the improvement using the combination of thesauri from the entire collection. The results indicate that despite only half the collection being used to build thesauri, the performance is not too degraded. In fact, we still get a significant improvement over the original queries. This means that the combined set of thesauri constructed from the sample collection is still useful for the entire collection.

5.5. Expansion using only thesauri intersection

We compare the retrieval results of using our method described above with the results of

Table 5
Example expansion term and their source thesauri

Source thesauri	Expansion terms
WordNet	Force
Co-occurrence	Uup, RUC, Downing
Pred-arg	Anglo-Irish
WordNet and co-occurrence	Treaty, paramilitary
WordNet and pred-arg	Irish, British, declaration
Co-occurrence and pred-arg	Sinn, Fein, Gerry, Adams, Heathrow, John, Major, Albert, Reynolds, Crossmaglen
All	IRA, protestant, catholic, politics, meeting, bomb, terrorism, Dublin, street, ceasefire, firebomb, gun, incident, unionist, constabulary, loyalist, crime, campaign, killing, arm, fire, explosion

expansion using only words that appear in the intersection of the suggestions made by all three thesauri. Table 7 shows the non-interpolated average precision of SMART with the *lnc.ltc* weighting method without expansion, our method, and expansion using only thesauri intersection. As we can see in Table 7, our method gives significantly better results than using only the intersection of the thesauri.

5.6. Comparison with relevance-feedback

We compare our method with query expansion using the relevance feedback technique, in which 30 documents among the documents retrieved in the initial retrieval are used for feedback. We use the Rocchio formula for term reweighting as follows:

$$Q_{\text{new}} = \alpha \cdot Q_{\text{old}} + \beta \sum_{r=1}^{n_{\text{rel}}} \frac{D_r}{n_{\text{rel}}} - \gamma \sum_{n=1}^{n_{\text{nonrel}}} \frac{D_n}{n_{\text{nonrel}}}$$

where α, β and γ are constants, D_r is the vector of a relevant document d_r, D_n is the vector of an irrelevant document d_n, n_{rel} is the number of relevant documents retrieved, and n_{nonrel} is the number of irrelevant documents. We set $\alpha = 8$, $\beta = 16$, and $\gamma = 4$ for this experiment (Buckley & Salton, 1994, 1995).

For ideal relevance feedback, 30 relevant documents are determined by referring to the relevance judgement on the test collection, and for the pseudo relevance feedback, the 30 top-ranked documents in the initial retrieval are assumed relevant.

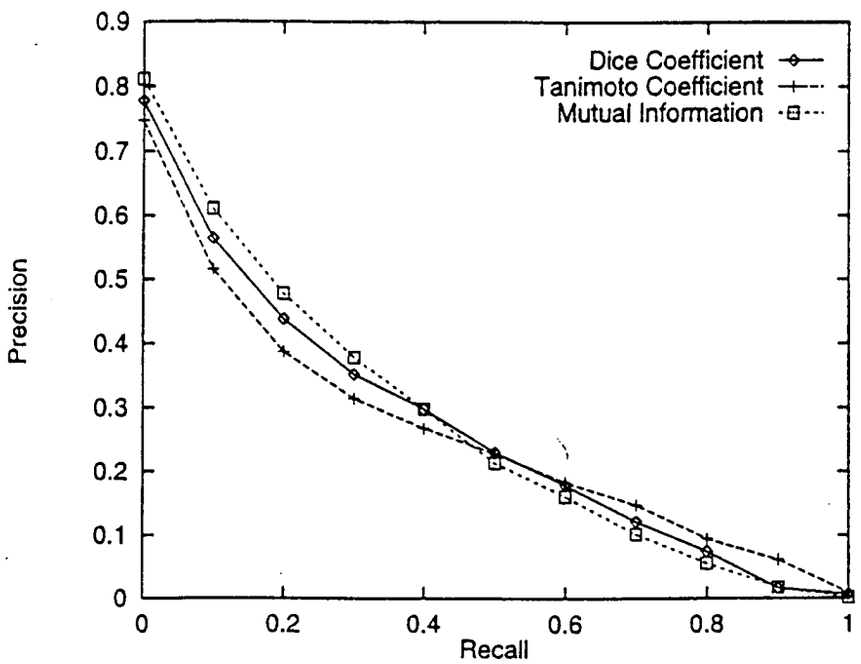


Fig. 3. Eleven-point precision for different similarity measures.

Table 6
Comparison of different similarity measures for different topic sections

Topic section	Base	Dice coefficient	MI	Tanimoto coefficient
Title	0.1175	0.2213 (+88.3%)	0.2314 (+96.9%)	0.2011 (+71.7%)
Description	0.1428	0.2590 (+81.4%)	0.2645 (+85.2%)	0.2492 (+75.3%)
All	0.1976	0.2573 (+30.2%)	0.2724 (+37.8%)	0.2654 (+34.5%)

Figs. 8–10 show 11-point recall-precision curves comparing our method, ideal relevance feedback, and pseudo relevance feedback using the title only, the description only, and all parts of the topic, respectively. It shows that our method significantly outperform pseudo relevance feedback and is slightly worse than ideal relevance feedback.

6. Discussion

The key techniques used in our method can be summarized as follows:

- Broadening thesaurus coverage by combining several thesauri.
- Weighting expansion terms to avoid wrong expansion.

The three types of thesauri used have different characteristics. Automatically constructed

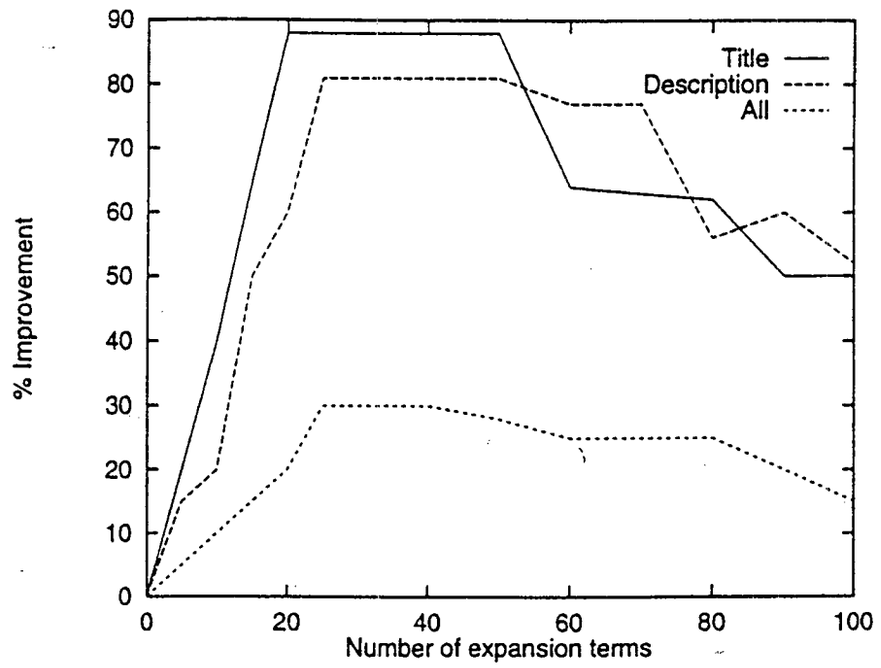


Fig. 4. Relation between the number of expansion terms and improvement.

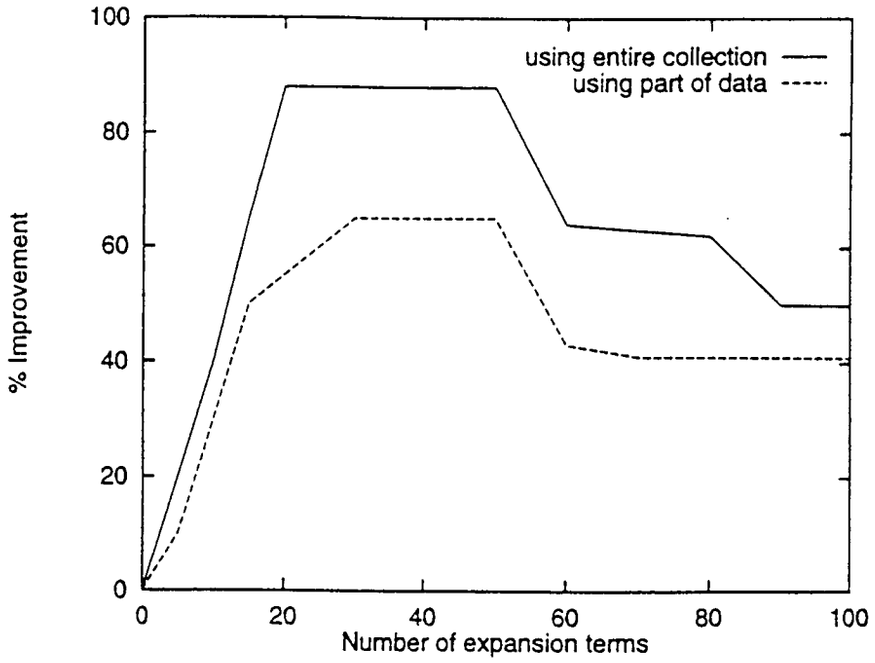


Fig. 5. Relation between the proportion of the corpus used and relative system improvement (title).

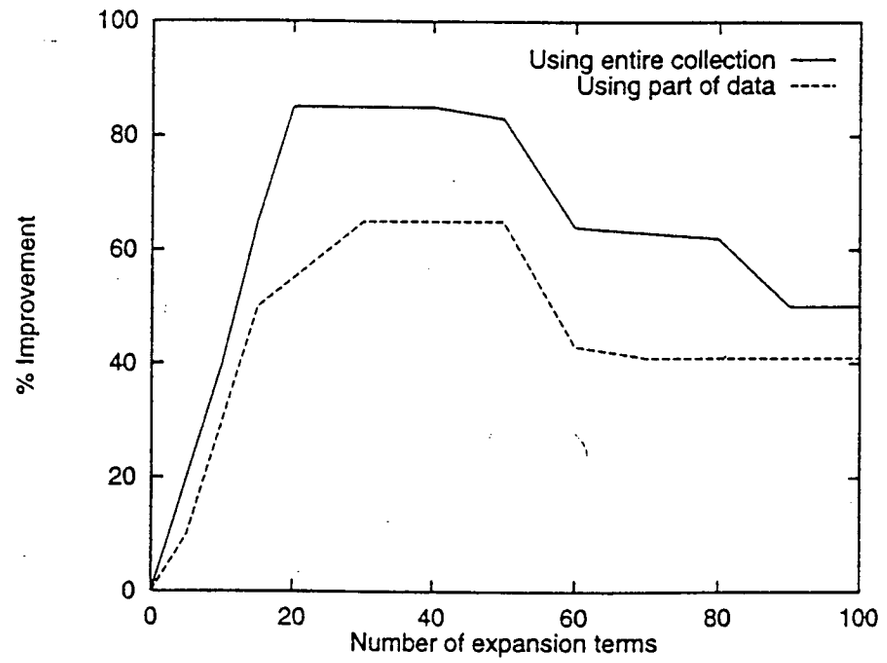


Fig. 6. Relation between the proportion of the corpus used and relative system improvement (description).

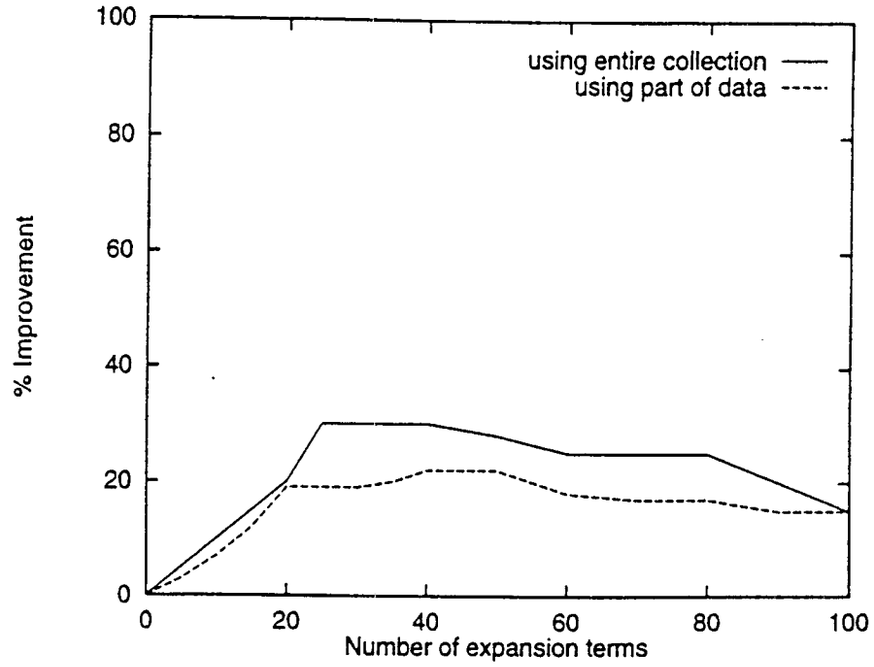


Fig. 7. Relation between the proportion of the corpus used and relative system improvement (all).

thesauri add not only new terms but also new relationships not found in WordNet. If two terms often co-occur in a document then those two terms are likely to bear some relationship. Why not only use the automatically constructed thesauri? The answer is that some general knowledge may be missing in automatically constructed thesauri.

The advantages of our weighting method can be summarized as follows:

- The weight of each expansion term considers the similarity of that term to all terms in the original query, rather than to just one query term.
- The weight of an expansion term also depends on its similarity in all types of thesauri.

This method can accommodate the polysemous word problem, because an expansion term taken from a different sense to the original query term sense is given very low weight. The reason for this is that the weighting method depends on all query terms and all of the thesauri. For example, the word "bank" has many senses in WordNet. Two such senses are a financial

Table 7
A comparison of the results of our method and using the intersection of the thesauri

Topic type	Without expansion	Our method	Intersection
Title	0.1175	0.2213 (+88.3%)	0.1553 (+32.1%)
Description	0.1428	0.2590 (+81.4%)	0.1789 (+25.3%)
All	0.1976	0.2573 (+30.2%)	0.2274 (+15.1%)

institution and river edge. In a document collection relating to financial banks, the river sense of "bank" will generally not be found in the co-occurrence based thesaurus because of a lack of documents talking about rivers. Even though (with small possibility) there may be some documents in the collection talking about rivers, if the query contained the finance sense of "bank" then the other terms in the query would also be concerned with finance and not rivers. Thus, rivers would only relate to the term "bank" and there would be no relationship with other terms in the original query, resulting in a low weight.

Our method only retrieves documents once and only needs extra time for computing the multiplication of original query terms weight and the similarity value between terms in the combined thesauri which are pre-computed, so that the time required by our method is proportional to the number of query terms. In contrast, relevance feedback needs to compute the similarity value between query terms and all documents in the second retrieval process, so that the time required by this method is proportional to the number of terms in all documents. Hence, our method requires less retrieval time than relevance feedback.

7. Conclusions and future works

We have proposed the use of heterogeneous thesauri for query expansion. The underlying idea is that each type of thesaurus has different characteristics and, therefore, their combination can provide a valuable resource for query expansion. Wrong expansion terms are

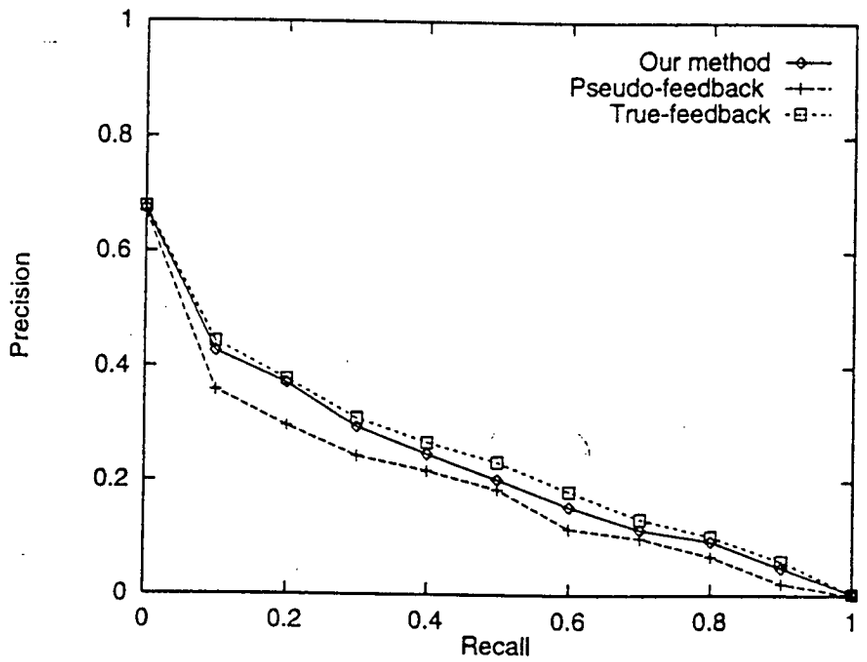


Fig. 8. Comparison of performance between our method and relevance feedback (title).

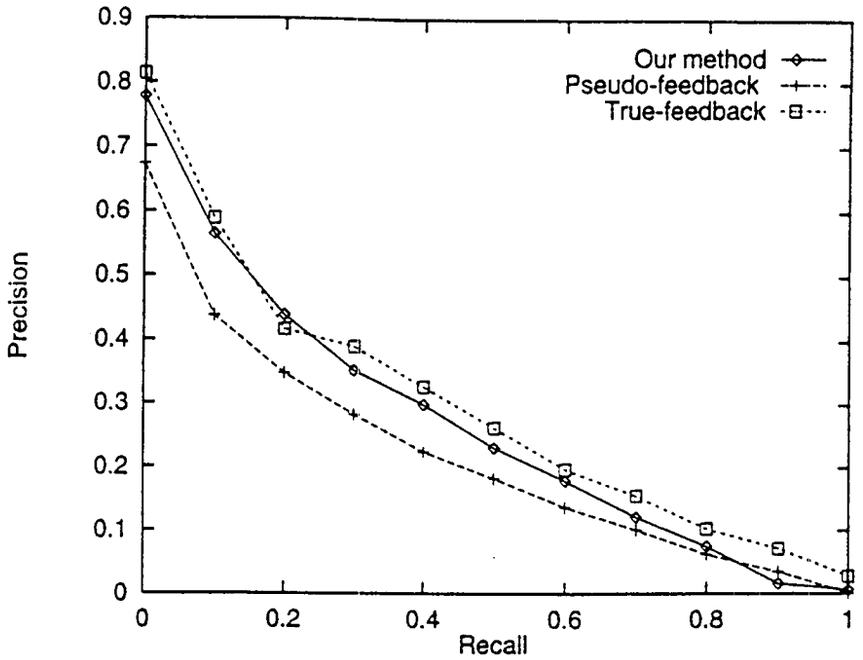


Fig. 9. Comparison of performance between our method and relevance feedback (description).

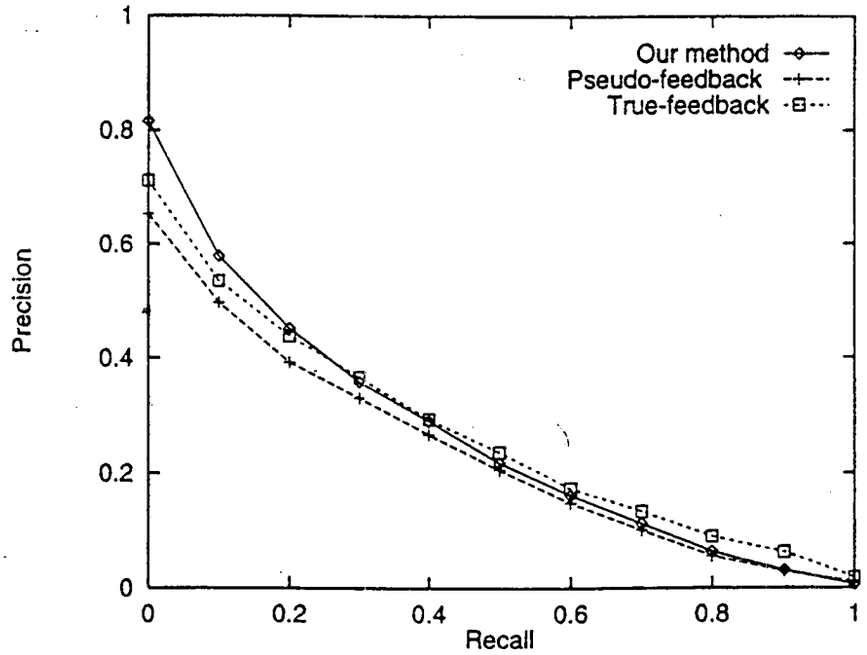


Fig. 10. Comparison of performance between our method and relevance feedback (all).

avoided by designing a weighting term method in which the weight of expansion terms not only depends on all query terms, but also on similarity measures in all types of thesauri. This is actually a kind of word sense disambiguation.

Experiments have shown that use of the combined set of thesauri gives better retrieval results than just using one type of thesaurus. Our method also out-performs pseudo relevance feedback, although more experiments are needed for further investigation.

Future research will include the use of parser with better performance and the use of anaphora resolution to accurately determine the nature of relationships involving proper names.

Acknowledgements

The authors would like to thank the anonymous referees and Timothy Baldwin for useful comments on the earlier version of this paper. We also thank Chris Buckley (SabIR Research) for support with SMART and Satoshi Sekine (New York University) for the Apple Pie Parser. This research is partially supported by JSPS project number JSPS-RFTF96P00502.

References

Buckley, C., & Salton, G. (1994). The effect of adding relevance information in a relevance feedback environment. In B. Croft, & C. van Rijsbergen, *Proceedings of the 17th Annual International ACM-SIGIR Conference on Research and Development in Information Retrieval, Dublin, Ireland* (pp. 292–300). London, UK: Springer-Verlag 3–6 July.

Buckley C., & Salton G. (1995). Automatic query expansion using SMART: TREC-3. In D. Harman (Ed.). *Overview of the Third Text Retrieval Conference (TREC-3)* (pp. 69–80). Gaithersburg, MD: NIST Special Publication 500-225.

Charniak, E. (1993). *Statistical Language Learning*. Cambridge, MA: MIT Press.

Chen, H., Schatz, B., Yim, T., & Fye, D. (1995). Automatic thesaurus generation for an electronic community system. *Journal of the American Society for Information Science*, *46*(3), 175–193.

Crouch, C. J. (1990). An approach to the automatic construction of global thesauri. *Information Processing and Management*, *26*(5), 629–640.

Crouch, C., & Yang, B. (1992). Experiments in automatic statistical thesaurus construction. In N. Belkin, P. Ingwersen, & A. M. Pesjtersen, *Proceedings of the 15th Annual International ACM-SIGIR Conference on Research and Development in Information Retrieval, Copenhagen, Denmark* (pp. 77–82). New York, NY: ACM Press 21–24 June.

Ekmekcioglu, F. (1992). Effectiveness of query expansion in ranked-output document retrieval systems. *Journal of Information Service*, *18*(2), 139–147.

Fox, E. A. (1980). Lexical relations enhancing effectiveness of information retrieval systems. *SIGIR Forum*, *15*(3), 6–36.

Grefenstette, G. (1992). Use of syntactic context to produce term association lists for text retrieval. In N. Belkin, P. Ingwersen, & A. M. Pesjtersen, *Proceedings of the 15th Annual International ACM-SIGIR Conference on Research and Development in Information Retrieval, Copenhagen, Denmark* (pp. 89–97). New York, NY: ACM Press 21–24 June.

Grefenstette, G. (1994). *Explorations in Automatic Thesaurus Discovery*. Dordrecht, Netherlands: Kluwer.

Hindle, D. (1990). Noun classification from predicate-argument structures. In R. C. Berwick, *Proceedings of the 28th*

Annual Meeting of the Association for Computational Linguistics, Pittsburgh, USA (pp. 268–275). Morristown, NJ: Association for Computational Linguistics 6–9 June.

Jing, Y., & Croft, B. (1994). An association thesaurus for information retrieval. In F. Brentano, & F. Seitz, *Proceedings of the RIAO'94 Conference, New York, USA* (pp. 146–160). Paris, France: C.I.D.-C.A.S.I.S 11–13 October.

Kristensen, J. (1993). Expanding end-users query statements for free text searching with a search-aid thesaurus. *Information Processing and Management, 29*(6), 733–744.

Miller, G., Beckwidth, R., Fellbaum, C., Gross, D., & Miller, K. (1990). Wordnet: an on-line lexical database. *International Journal of Lexicography: Special Issue, 3*(4), 235–244.

Paice, C. D. (1991). A thesaural model for information retrieval. *Information Processing and Management, 27*(5), 433–447.

Peat, J., & Willett, P. (1991). The limitations of term co-occurrence data for query expansion in document retrieval systems. *Journal of the American Society for Information Science, 42*(5), 378–383.

Qiu, Y., & Frei, H. (1993). Concept based query expansion. In R. Korfhage, E. Rasmussen, & P. Willett, *Proceedings of the 16th Annual International ACM-SIGIR Conference on Research and Development in Information Retrieval, Pittsburgh, USA* (pp. 160–169). New York, NY: ACM Press 27 June–1 July.

Resnik, P. (1995). Using information content to evaluate semantic similarity in a taxonomy. In C. Mellish, *Proceedings of the 14th International Joint Conference on Artificial Intelligence (IJCAI-95), Montreal, Canada* (pp. 448–453). San Francisco, CA: Morgan Kaufmann 20–25 August.

Ruge, G. (1992). Experiments on linguistically-based term associations. *Information Processing and Management, 28*(3), 317–332.

Salton, G. (1971). *The SMART Retrieval System Experiments in Automatic Document Processing*. Englewood Cliffs, NJ: Prentice-Hall.

Salton, G., & Buckley, C. (1990). Improving retrieval performance by relevance feedback. *Journal of the American Society for Information Science, 41*(4), 288–297.

Salton, G., & McGill, M. (1983). *An Introduction to Modern Information Retrieval*. New York, NY: McGraw-Hill.

Schutze, H., & Pedersen, J. (1994). A cooccurrence-based thesaurus and two applications to information retrieval. In F. Brentano, & F. Seitz, *Proceedings of the RIAO'94 Conference, New York, USA* (pp. 266–274). Paris, France: C.I.D.-C.A.S.I.S 11–13 October.

Schutze, H., & Pederson, J. (1997). A cooccurrence-based thesaurus and two applications to information retrieval. *Information Processing and Management, 33*(3), 307–318.

Sekine, S., & Grishman, R. (1995). A corpus-based probabilistic grammar with only two non-terminals. In *Proceedings of the International Workshop on Parsing Technologies, Prague, Czech Republic.*

Smeaton, A., Kelledy, F., & O'Donnel, R. (1996). TREC-4 Experiments at Dublin City University: thresholding postings lists, query expansion by word–word distances and POS tagging of Spanish text. In D. Harman, *Proceedings of the Fourth Text Retrieval Conference (TREC-4)* (pp. 373–390) NIST Special Publication 500-236, Gaithersburg, MD.

Voorhees, E. (1993). Using WordNet to disambiguate word senses for text retrieval. In R. Korfhage, E. Rasmussen, & P. Willett, *Proceedings of the 16th Annual International ACM-SIGIR Conference on Research and Development in Information Retrieval, Pittsburgh, USA* (pp. 171–180). New York, NY: ACM Press 27 June–1 July.

Voorhees, E. (1994). Query expansion using lexical-semantic relations. In B. Croft, & C. van Rijsbergen, *Proceedings of the 17th Annual International ACM-SIGIR Conference on Research and Development in Information Retrieval, Dublin, Ireland* (pp. 61–69). London, UK: Springer-Verlag 3–6 July.

Voorhees, E., & Harman, D. (1999). Overview of the seventh text conference (TREC-7). In *Proceedings of the Seventh Text Retrieval Conference* NIST Special Publication 500-242.

Xu, J., & Croft, B. (1996). Query expansion using local and global document analysis. In H.-P. Frei, & P. Schauble, *Proceedings of the 19th Annual International ACM-SIGIR Conference on Research and Development in Information Retrieval, Zurich, Switzerland* (pp. 4–11). New York, NY: ACM Press 18–22 August.

ARTICLE IN PRESS

ELSEVIER

Information Processing and Management xxx (2004) xxx–xxx

INFORMATION PROCESSING & MANAGEMENT

www.elsevier.com/locate/infoproman

Reformulation of queries using similarity thesauri

Ángel F. Zazo *, Carlos G. Figuerola, José L. Alonso Berrocal, Emilio Rodríguez

University of Salamanca, REINA Research Group, Cl Fco. de Vitoria 6-16, 37008 Salamanca, Spain

Received 23 September 2003; accepted 14 May 2004

Abstract

One of the major problems in information retrieval is the formulation of queries on the part of the user. This entails specifying a set of words or terms that express their informational need. However, it is well-known that two people can assign different terms to refer to the same concepts. The techniques that attempt to reduce this problem as much as possible generally start from a first search, and then study how the initial query can be modified to obtain better results. In general, the construction of the new query involves expanding the terms of the initial query and recalculating the importance of each term in the expanded query. Depending on the technique used to formulate the new query several strategies are distinguished. These strategies are based on the idea that if two terms are similar (with respect to any criterion), the documents in which both terms appear frequently will also be related. The technique we used in this study is known as query expansion using similarity thesauri.
© 2004 Published by Elsevier Ltd.

Keywords: Information retrieval; Query expansion; Similarity thesaurus; Term weighting

1. Introduction

In Information Retrieval (IR) one of the main problems facing users is that of expressing their informational need in a query suitable for the retrieval system. Apart from the requirements of the system for formulating the query, the main problem lies in determining the set of words or terms that express this need semantically. The problem is aggravated owing to the effect of inconsistency in the subjective assigning of

* Corresponding author.
E-mail address: figue@usal.es (Á.F. Zazo).
URL: http://REINA.usal.es.

0306-4573/$ - see front matter © 2004 Published by Elsevier Ltd.
doi:10.1016/j.ipm.2004.05.006

terms to concepts, or what is the same, that two people use different words for defining the same concepts (Furnas, Landauer, Gomez, & Dumais, 1987).

In current IR systems with complete text this approach is especially important. The content of a document is represented using the words that appear in it. The set of index terms is formed by all the words in the document collection after pre-processing the text, which usually includes the elimination of stop words, stemming, selection of terms by semantic category, etc. The user's informational need must also be represented using these terms. The main drawback to this approach is the well-known *vocabulary mismatch problem*. On the one hand, there are different words that express the same concepts (synonymy) and on the other hand, the same word can have several meanings (polysemy). In these systems it is assumed that two documents are similar if they contain the same words. But this is the problem: the same concept can be expressed with different words, and one same word can appear in documents dealing with completely different subjects.

In this situation, depending on their experience, users will normally have to reformulate their queries until they obtain suitable results. The user begins by sending an initial query to the retrieval system which returns a set of documents ordered according to some relevance criterion of the system. After examining the documents retrieved, the user has to reformulate the query to obtain more relevant results. The greater the number of terms in the query, the smaller the problem, since it will surely include a greater number of index terms that represent relevant documents (Xu & Croft, 2000). In general, the construction of the new query involves expanding terms of the initial query and recalculating the importance of each term in the expanded query (*reweighting of terms*).

In this study we focus on short queries, i.e. queries with one or only a few terms. These queries are particularly interesting because they are the typical ones made on the search engines of the Internet. It is well-known that queries on the web tend to be very short, between one and three terms per query (Jansen, Spink, & Saracevic, 2000; Wolfram, Spink, Janses, & Saracevic, 2001). Retrieval results can be improved by expanding the initial query, i.e., including better terms in the query that provide more relevant documents than the original query. The problem lies in finding these better terms.

2. Query expansion

To expand the query, words or phrases with a similar meaning to those in the initial query must be used. A dictionary or general thesaurus could be used in this process. A thesaurus is a classification system compiled of words and/or phrases organised with the objective of facilitating the expression of ideas. They have been used in retrieval information in the process of creating or expanding queries, i.e. in the process of expressing or broadening the informational needs of a user in a query. Basically this is done by selecting the terms closest to what the user wants. Unfortunately, the use of a general thesaurus for formulating the user's need does not give good results (Voorhees, 1994), mainly because the relations in a general thesaurus are not valid in the local context of user and document collection. Better results are obtained if thesauri or query expansion techniques constructed from the document collection on which the search is launched are used. In any case, reformulation of the query involves two major problems: the choice of the most suitable terms and the weighting of the new terms. This process can be done automatically, i.e., without the intervention of the user, or with the user's collaboration. In the latter case, the user is offered a graphic interface with the results of the search. In Belkin et al. (2001) the evolution of research in the reformulation of queries in the interactive task of the famous TREC (*Text REtrieval Conference*) is analysed and some user interfaces can be seen. With this interface the users can visualise and select the documents and terms that they consider more relevant to their initial search and the system will use them to reformulate the query.

Mainly three mechanisms are distinguished, depending on the technique used for formulating the new query (Baeza-Yates & Ribeiro-Neto, 1999): (i) the so-called query feedback with the intervention of the user (*user relevance feedback*); (ii) query reformulation, using exclusively information from the documents retrieved (*local analysis*) and (iii) reformulation using overall information from the whole document collection (*global analysis*). These strategies are based on the idea that if two items are similar (with respect to any criterion) the terms that frequently co-occur in them will be related. Based on this idea, semantic relations between the terms are sought. The *Association Hypothesis* (van Rijsbergen, 1979, p. 104) says:

> If an index term is good at discriminating relevant from non-relevant documents then any closely associated index term is also likely to be good at this.

Indeed, considering that the terms that appear in the query are good at discriminating relevant and non-relevant documents, the related terms will also be so. This would allow us to add them to the original query.

The first of the techniques indicated is user relevance feedback (RF). This is a mechanism that has been the object of much study and one of those that give the best results. In RF process the user marks the documents retrieved as relevant and non-relevant, and the system reformulates the query with the terms of these documents. The main advantage is that it is a process guided by the user's criteria of relevance, which is the basis for trying to increase the importance of certain terms and decrease that of others (Salton & Buckley, 1990). One of the algorithms most used is that of Rocchio (Rocchio, 1971). Our research group has also made some experiments on the subject (Figuerola, Zazo, & Berrocal, 2002). A similar technique that does not require the presence of the user is pseudo relevance feedback (PRF). In PRF the first 10 or 20 retrieved documents assume relevant and the algorithm of Rocchio is applied to obtain the reformulated query. It is a simple local expansion method that obtains good results (Mitra, Singhal, & Buckley, 1998).

The other two techniques for query expansion look for relations or associations between terms: synonyms, derivatives (inflectional and derivational variants) or proximal ones (which have a minimum distance in number of words). Clustering algorithms are normally used for this. These algorithms are sometimes based on counting the co-occurrences of terms in the documents, and other times on finding similarities under other criteria. Global analysis looks for relations between the terms of the whole document collection, whereas local analysis only considers the documents retrieved and shown to the user for one particular query. Both seek to construct a matrix or thesaurus of relations between the terms, in a global or local sense, and use it to expand the original query with the terms considered best related. Two important strategies stand out in local analysis: on the one hand, what is known as *local clustering*, based on the study by Attar and Fraenkel (1977), and, on the other hand, the combination of certain techniques of local and global analysis, called *local contextual analysis* (Xu & Croft, 1996, 2000). Both obtain good results based on the documents retrieved.

The global analysis techniques extract information from the whole collection of documents and use it to expand the query. Research in this field has been developing since before the 1960s. Most of it has had as a basis the study of co-occurrences of terms. It was not until the beginning of the 1990's, however, that satisfactory results were obtained (Peat & Willet, 1991), mainly using clustering of term, although also of documents. Term clustering is used for the construction of thesauri. Based on the criteria used in its construction, some approaches are distinguished: thesauri constructed using simple measurement of term co-occurrences (Minker, Wilson, & Zimmerman, 1972); using clustering of documents to create a thesaurus of infrequent terms (Crouch, 1990); thesauri constructed making the transposition of the document-term matrix (similarity thesauri) (Qiu & Frei, 1993); thesauri constructed from the association of terms and phrases (*phrase-finder*) (Jing & Croft, 1994); thesauri based on syntactic information (Grefenstette, 1992a); and thesauri constructed from several sources of expansion (WordNet, general dictionaries, local thesauri, manual lexicon or parallel corpus) (Mandala, Tokunaga, & Tanaka, 1999; Xu, Fraser, & Weischedel, 2001).

ARTICLE IN PRESS

The main strategies carried out, both automatically and with the intervention of the user, have been based on the simple use of co-occurrence values, the classification of documents or the use of syntactical contexts, with results highly dependent on the document collection to which they were applied, or on parameters that are difficult to calculate or to adjust (Crouch & Yang, 1992; Grefenstette, 1992b; Han, Fujii, & Croft, 1995; Jing & Croft, 1994; Mandala, Tokunaga, & Tanaka, 2000; Qiu & Frei, 1993; Schutze, Hull, & Pedersen, 1995; Xu et al., 2001). In contrast with these methods there is another that is attempting to create a similarity global thesaurus as a basis for query expansion.

3. Similarity thesauri

The target here is to construct a similarity thesaurus that makes it possible to expand the complete query (*query concept*), and not only each individual term separately (Qiu & Frei, 1993). A similarity thesaurus is a matrix that is constructed using relations between terms. In its preparation the co-occurrence of terms in the documents is not measured, instead a mechanism is used by which each term in the collection is characterised by the documents in which it appears. This turns the classic concept of information retrieval systems upside down (in these, documents are characterised by index terms, i.e., the terms are used to represent the documents). To construct the similarity thesaurus, the terms of the collection are considered documents, and the documents are used as index terms (they are usually called *index documents*); in other words, the documents can be considered capable of representing the terms.

In order to apply this approach we shall take as a basis the vector space model. In the vector space model (Salton, 1968) each document d_i in the collection of N documents is represented by a vector $\vec{d}_i = (w_{i1}, w_{i2}, \ldots, w_{im})$, where w_{ij} indicates the weight of the index term t_j in document d_i, and m is the number of index terms in the collection. The weights are normally calculated using the *tf-idf* scheme (Salton & Yang, 1973). The query is also represented in the vector space of terms using a vector, $\vec{q} = (q_1, q_2, \ldots, q_m)$. Each element q_j expresses the degree to which the term t_j represents the informational needs of the person making the query. The main objective of an information retrieval system is to provide information appropriate to the query made by the user. This process depends enormously on the weighting scheme of the terms and subsequent calculation of similarity made by the system. In Salton and Buckley (1988) 287 different combinations of assigning weights to terms of documents and queries were experimented with, and we have followed the indications that appear there in order to obtain the best results. The easiest way of obtaining the degree of similarity between a query and a document is to calculate the scalar product of the vectors that represent them (van Rijsbergen, 1979). This is a simple process and one of the most used. So that the similarity value will be between 0 and 1, the vectors of the documents and queries are normalized.

Below we shall see how the similarity thesaurus can be constructed in the vector space model. The idea is to turn around the representation model described and instead of considering the documents as represented by index terms, we consider that the index terms can be defined using the documents (*index documents*). Turning the vector model around, each term t_i in the collection of m terms will be represented by a vector of N components in the vector space of documents, $\vec{t}_i = (p_{i1}, p_{i2}, \ldots, p_{iN})$, p_{ij} being a value that expresses the weight of index document d_j in the representation of the term t_i. In order to compute the value of p_{ij} the *tf-idf* scheme is used, but the roles of terms and documents is inverted. The weights are normalized in order to have unit vectors. The calculation proposed in Qiu and Frei (1993) is the one indicated in (1). This calculation derives from the inversion of the *ann* scheme (in SMART's term weight triple notation (Salton & Buckley, 1988)) for documents, plus *itf* factor, see below. This is the one we used in our experiments.

$$p_{ij} = \frac{\left(0.5 + 0.5\frac{f_{ij}}{\max_k(f_{ik})}\right) \cdot itf_j}{\sqrt{\sum_{u=1}^{N}\left(0.5 + 0.5\frac{f_{iu}}{\max_k(f_{ik})}\right)^2 \cdot itf_u^2}} \tag{1}$$

where f_{ij} is the number of times that the term t_i appears in document d_j, $\max_k(f_{ik})$ is the maximum of the frequency values for the term t_i in the whole document collection (i.e., it will be the value f_{ik}, d_k being the document where it appears most), $itf_j = \log\frac{n}{|d_j|}$ is the *inverse term frequency* for document d_j, $|d_j|$ being the number of different terms there are in that document.

Calculation of the *inverse term frequency* shows that a short document plays a more important role than a long one. If two terms co-occur in a long document, the probability of them being similar is smaller than if they co-occur in a short one. To calculate the similarity between two terms, t_i and t_j, the scalar product is also used.

$$\mathrm{SIM}(t_i, t_j) = \vec{t}_i^{\mathrm{T}} * \vec{t}_j = \sum_{k=1}^{N} p_{ik} \cdot p_{jk} \tag{2}$$

Calculation for all the terms pairs in the collection produces the similarity thesaurus. It is a symmetric matrix, with values between 0 and 1. Its construction is computationally costly, although it is only done once. If documents are added to the collection the values for the terms included in the new documents must be updated. Once again we must point out that the thesaurus is constructed taking into consideration the representation of terms using index documents, and not only co-occurrence information. The importance is the weight of each document in the representation of terms.

3.1. Query expansion

The objective when using the similarity thesaurus is to expand the whole query, and not just each individual term separately. For a term to be able to be added to the query there must be a high similarity between this term and all the terms in the query. In the vector space of the terms query q is represented by the vector $\vec{q} = (q_1, q_2, \ldots, q_m)$, where q_i is the weight of term t_i in the query. However, given that the similarity thesaurus has been constructed using the vector space of documents, we must transfer that vector to this space so as to be able to compute the similarity between the whole query and all the index terms in the collection. The term t_i is given by the vector $\vec{t}_i$ in the document space, so that for the whole query, in this space, it will have an importance of $q_i \cdot \vec{t}_i$. We consider that in this space the query only depends on the terms included in it.

Once the query has been represented in the document space, we must obtain the terms most similar to it. We use the measurement of the scalar product with each of the terms t in the collection.

$$\mathrm{sim}(q, t) = \vec{q}^{\mathrm{T}} * \vec{t} = \left(\sum_{t_i \in q} q_i \cdot \vec{t}_i\right)^{\mathrm{T}} * \vec{t} = \sum_{t_i \in q} q_i \cdot (\vec{t}_i^{\mathrm{T}} * \vec{t}) = \sum_{t_i \in q} q_i \cdot \mathrm{SIM}(t_i, t) \tag{3}$$

The similarity values between terms are the entries in the similarity thesaurus which have already been calculated. Thus, all the terms of the collection can be put in order according to the value obtained from (3). In general not all the terms are expanded, but only those which have the highest similarity values. We shall denote as r the number of terms that will be added to the original query. The weight, in the space of terms, associated with each term t_e, which will be added to the query, remains to be determined. It seems natural to consider it according to the similarity:

Table 1
Characteristics of the Collection EFE'94

No. of documents	215,738
No. of queries	50
No. of index terms	352,777
Mean number of words per document	333.68 (max 2210, min 9)
Mean number of unique index terms per document	120.48

$$q_e = \frac{\mathrm{sim}(q, t_e)}{\sum_{t_i \in q} q_i} \tag{4}$$

The value of the weight of each expanded term is that given in (4). That is, new terms are added to the initial query and the weight of the already existing terms can be modified if they appear among the first r selected for expansion.

4. Experiments

In our experiments we employed the test collection used in several CLEF [1] studies (Peters & Braschler, 2001; Zazo, Figuerola, Berrocal, & Gómez, 2002). The characteristics of the collection are given in Table 1. The mean number of terms was considered after eliminating stop words. The document collection came from the news agency EFE, from all the news items of 1994: 215,718 documents (513 MB of information), stored in files, one for each day of 1994. Each file contains several documents, and in these there are fields delimited with SGML labels, as indicated in the example in Fig. 1. The relevant fields are `TITLE` and `TEXT`, since the rest contain information such as date, section of the newspaper, etc. The mean number of words per document, considering these fields, is 333.68.

Together with the documents there is a test set of 50 queries in Spanish. Each query is divided into three fields, `ES-title`, `Es-desc` (description) and `ES-narr` (narrative), as can be seen in Fig. 2. Queries can be posed with one, several or all the fields.

In our experiments we considered an index term to be composed of any set of alphanumerical characters, i.e., we also included numbers as index terms. In the lexical pre-processing of the text the terms of the documents and queries that can be used as index terms are obtained. This pre-processing was carried out in several stages. In the first we eliminated all the non-alphanumerical characters, did not detect proper names or acronyms and stored the index terms in small letters without considering accents.

The second action performed in pre-processing the text was the elimination of stop words. With the objective of reducing the number of index terms, words that are not very significant in the information retrieval process because of their slight semantic capacity or high frequency are not included. This set of words, known as stop words, is composed of prepositions, articles, adverbs, conjunctions, possessives, demonstratives, pronouns, some verbs and some nouns. With the elimination of stop words, the aim is to reduce the noise that they could introduce in the retrieval. To eliminate stop words, the text was separated into words, which were then checked against the stop word list. In our experiments we used a list of 573 words.

The next step consisted of stemming. Stemming is the process whereby morphological variations of the terms are sought in order to extract the common root. The canonical form of the root represents the variations of the terms deriving from it. In our case, we did not apply stemming.

[1] CLEF collections are property of CLEF Consortium and ELDA.

<DOC>

<DOCNO>EFE19940101-00002</DOCNO>

<DOCID>EFE19940101-00002</DOCID>

<DATE>19940101</DATE>

<TIME>00.34</TIME>

<SCATE>VAR</SCATE>

<FICHEROS>94F.JPG</FICHEROS>

<DESTINO>ICX MUN EXG</DESTINO>

<CATEGORY>VARIOS</CATEGORY>

<CLAVE>DP2404</CLAVE>

<NUM>100</NUM>

<PRIORIDAD>U</PRIORIDAD>

<TITLE> IBM-WATSON

— FALLECIO HIJO FUNDADOR EMPRESA DE COMPUTADORAS

</TITLE>

<TEXT> Nueva York, 31 dic (EFE).- Thomas Watson junior, hijo del fundador de International Business Machines Corp. (IBM), falleció hoy, viernes, en un hospital del estado de Connecticut a los 79 años de edad, informó un portavoz de la empresa.

Watson falleció en el hospital Greenwich a consecuencia de complicaciones tras sufrir un ataque cardíaco, añadió la fuente.

El difunto heredó de su padre una empresa dedicada principalmente a la fabricación de máquinas de escribir y la transformó en una compañía líder e innovadora en el mercado de las computadoras. EFE

PD/FMR

01/01/00-34/94

</TEXT>

</DOC>

Fig. 1. A document from the collection.

```
<top>
<num> C042 </num>
<ES-title> Naciones Unidas y Estados Unidos invaden Haití</ES-title>
<ES-desc> Encontrar documentos sobre la invasión de Haití por los
soldados de la ONU y de los Estados Unidos. </ES-desc>
<ES-narr> Los documentos comentan tanto la discusión sobre la
decisión de la ONU de enviar las tropas americanas a Haití, como la
invasión misma. Se habla también de sus consecuencias
directas.</ES-narr>
</top>
```

Fig. 2. A query from the collection.

Next, we selected the terms or groups of terms that would be the index terms. This is normally done according to the syntactical nature of the term, since those that act grammatically as nouns usually have a greater semantic content than verbs, adjectives or adverbs. Our selection of terms was not made on the basis of any syntactical criterion. Simply, those remaining after eliminating the stop words were considered as index terms.

The last step in the pre-processing of terms was the construction or application of a thesaurus that allowed the query to be expanded. Query expansion has already been described. The objective of this study was to show the characteristics of query expansion using similarity thesauri on the EFE'94 document collection.

Once the index terms of the test collection had been obtained, representation of documents and queries was done using the *tf-idf* weighting mechanism and the recommendations of Salton and Buckley (1988). We used the scalar product to calculate similarities between documents and queries. In the expansion of each query a local similarity thesaurus was calculated for the terms included in it. The related terms as described in Section 3 were obtained and arranged in decreasing order to then take the first r. The results were evaluated by calculating the average precision for all the queries in the collection in three values representative of recall: a low value of 25%, a medium value of 50% and a high value of 75%. Then the mean of these three values was taken.

In this study we also wished to measure the influence of the number of terms in the original query on the results. For this purpose we did several tests considering different fields of the query: test T (`ES-title` field), test D (`ES-desc` field) and test N (`ES-narr` field). Table 2 shows the improvement in query expan-

Table 2
Mean average precision on 50 queries

Tests	T	D	N
Mean no. of terms per query	2.74	8.40	16.66
Average precision original queries	0.3218	0.3717	0.3634
Average precision expanded queries with $r = 500$	0.3505	0.3902	0.3759
Improvement	8.92%	4.98%	3.44%

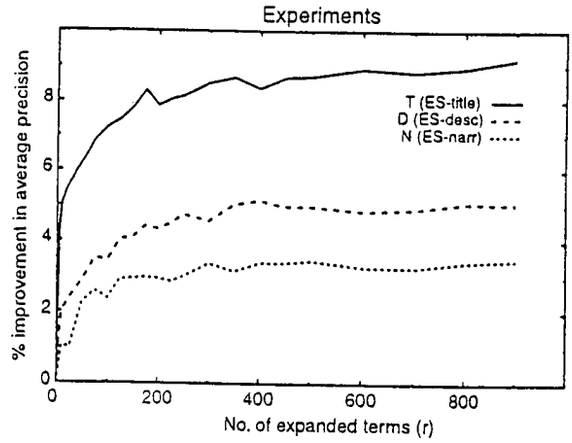


Fig. 3. Improvement in the average precision according to the number of terms expanded. Tests: T (`Es-title` field), D (`Es-desc` field) and N (`Es-narr` field).

sion, with an expansion of 500 terms added to the original query. It can be noted that the greater the number of terms in the original query, the smaller the improvement. This was to be expected, since long queries contain greater description of the user's informational need than short queries.

Fig. 3 shows the results of our experiments according to the number of terms expanded. We can see that the improvement follows the same course for all three tests. A major improvement is noted with the expansion of a few terms, but after a certain number (around 250) the improvement increases very slowly, and continues to increase, perhaps indefinitely. This seems to coincide with the experiments of other authors (Qiu & Frei, 1993) when the test collection is fairly large, as is our case. This may have to do with the fact that more discriminating terms are needed in large test collections.

5. Conclusions

The technique described obtains good results in query expansion. A thesaurus of similarity between terms was constructed, taking advantage of the possibilities the documents have to represent the terms. This duality in the representation of information allowed us to construct the thesaurus taking into account a weighting scheme according to the length of the documents and not just the number of times that a term appears in that document. The main characteristic resides, however, in the fact that the expanded terms were chosen and weighted taking into consideration the terms of the whole query, and not each individual term separately.

Moreover, as opposed to other techniques of local analysis that base their expansion on the terms of a few documents with user relevance feedback, here we used a technique that related all the terms in the test collection to each other. In fact, our results are worse than obtained with user relevance feedback, mainly because the expansion we used did not incorporate any type of additional relevance information, which other methods do incorporate. It is, therefore, a mechanism that can be executed automatically a some more characteristic of the information retrieval system. The main drawback lies in the high computational cost required for the construction of the similarity thesaurus. For new documents the task is more simple, since the values of the terms appearing in them are modified.

One very interesting aspect is that the smaller the query, the more it benefits from expansion. Considering that the queries on the search engines on the Internet usually contain one, two or three terms, this

technique can be especially useful in this context. However, considering that the volatility of the information on the Internet is very high, and that major modification of the thesaurus would be required, since the number of documents that would disappear from the collection and the number that would be incorporated is enormous, we believe that this technique can only be used in more static situations.

References

Attar, R., & Fraenkel, A. S. (1977). Local feedback in full-text retrieval system. *Journal of the ACM, 24*(3), 397–417.

Baeza-Yates, R., & Ribeiro-Neto, B. (1999). *Modern information retrieval.* Harlow, England: Addison-Wesley.

Belkin, N. J., Cool, C., Kelly, D., Lin, S.-J., Park, S. Y., Pérez-Carballo, J., & Sikora, C. (2001). Iterative exploration, design and evaluation of support for query reformulations in interactive information retrieval. *Information Processing & Management, 37*(3), 403–434.

Crouch, C. J. (1990). An approach to the automatic construction of global thesauri. *Information Processing & Management, 26*(5), 629–640.

Crouch, C. J., & Yang, B. (1992). Experiments in automatic statistical thesaurus construction. In N. J. Belkin, P. Ingwersen, & A. M. Pejtersen (Eds.), *Proceedings of the 15th annual international ACM-SIGIR conference on research and development in information retrieval, Copenhagen, Denmark, June 21–24* (pp. 77–88). ACM Press.

Figuerola, C. G., Zazo, Á. F., & Berrocal, J. L. A. (2002). La interacción con el usuario en los sistemas de recuperación de información: Realimentación por relevancia. *Scire, 8*(1), 87–94.

Furnas, G. W., Landauer, T. K., Gomez, L. M., & Dumais, S. T. (1987). The vocabulary problem in human–system communication. *Communications of the ACM, 30*(11), 964–971.

Grefenstette, G. (1992a). Sextant: exploring unexplored contexts for semantic extraction from syntactic analysis. In *30th annual meeting of the association for computational linguistics, Newark, Delaware, 28 June–2 July. ACL'92* (pp. 324–326).

Grefenstette, G. (1992b). Use of syntactic context to produce term association lists for text retrieval. In N. J. Belkin, P. Ingwersen, & A. M. Pejtersen (Eds.), *Proceedings of the 15th annual international ACM-SIGIR conference on research and development in information retrieval, Copenhagen, Denmark, June 21–24* (pp. 89–97). ACM Press.

Han, C., Fujii, H., & Croft, W. B. (1995). *Automatic query expansion for Japanese text retrieval.* Technical Report UM-CS-1995-011, Department of Computer Science, Lederle Graduate Research Center, University of Massachusetts. Available: ftp://ftp.cs.umass.edu/pub/techrept/techreport/1995/UM-CS-1995-011.ps.

Jansen, B. J., Spink, A., & Saracevic, T. (2000). Real life, real users, and real needs: a study and analysis of user queries on the web. *Information Procession & Management, 36*(2), 207–227.

Jing, Y., & Croft, W. B. (1994). An association thesaurus for information retrieval. In *Proceedings of RIAO-94, 4th international conference "Recherche d'Information Assistee par Ordinateur", New York, US* (pp. 146–160).

Mandala, R., Tokunaga, T., & Tanaka, H. (1999). Combining multiple evidence from different types of thesaurus for query expansion. In M. Hearst, F. Gey, & R. Tong (Eds.), *Proceedings of the 22nd annual international ACM-SIGIR conference on research and development in information retrieval, University of Berkeley, California, USA, August, 1999* (pp. 191–197). ACM.

Mandala, R., Tokunaga, T., & Tanaka, H. (2000). Query expansion using heterogeneous thesauri. *Information Processing & Management, 36*(3), 361–378.

Minker, J., Wilson, G. A., & Zimmerman, B. H. (1972). An evaluation of query expansion by the addition of clustered terms for a document retrieval system. *Information Storage and Retrieval, 8*(6), 329–348.

Mitra, M., Singhal, A., & Buckley, C. (1998). Improving automatic query expansion. In *Proceedings of the 21st annual international ACM-SIGIR conference on research and development in information retrieval, Melbourne, Australia, August 24–28, 1998* (pp. 206–214). ACM.

Peat, H. J., & Willet, P. (1991). The limitations of term co-occurrence data for query expansion in document retrieval systems. *Journal of the American Society for Information Science, 42*(5), 378–383.

Peters, C., & Braschler, M. (2001). European research letters: cross-language system evaluation: the CLEF campaigns. *Journal of the American Society for Information Science and Technology, 53*(12), 1067–1072.

Qiu, Y., & Frei, H.-P. (1993). Concept-based query expansion. In R. Korfhage, E. M. Rasmussen, & P. Willet (Eds.), *Proceedings of the 16th annual international ACM-SIGIR conference on research and development in information retrieval, Pittsburgh, PA, USA, June 27–July 1, 1993* (pp. 160–169). ACM.

Rocchio, J. J. (1971). Relevance feedback in information retrieval. In G. Salton (Ed.), *The SMART retrieval system. Experiments in automatic document processing* (pp. 313–323). Englewoods Cliffs, NJ: Prentice Hall.

Salton, G. (1968). *Automatic information organization and retrieval.* New York: McGraw-Hill.

Salton, G., & Buckley, C. (1988). Term-weighting approaches in automatic text retrieval. *Information Processing & Management, 24*(5), 513–523.

ARTICLE IN PRESS

Salton, G., & Buckley, C. (1990). Improving retrieval performance by relevance feedback. *Journal of the American Society for Information Science, 41*(4), 288–297.

Salton, G., & Yang, C. S. (1973). On the specification of term values in automatic indexing. *Journal of Documentation, 29*(4), 351–372.

Schutze, H., Hull, D. A., & Pedersen, J. O. (1995). A comparison of classifiers and document representations for the routing problem. In E. A. Fox, P. Ingwersen, & R. Fidel (Eds.), *Proceedings of the 18th annual international ACM SIGIR conference on research and development in information retrieval, Seattle, Washington, USA, July 9–13, 1995 (special issue of the SIGIR Forum)* (pp. 229–237). ACM.

van Rijsbergen, C. J. (1979). *Information Retrieval*. Department of Computer Science, University of Glasgow, second edition. Available: http://www.dcs.gla.ac.uk/Keith/.

Voorhees, E. M. (1994). Query expansion using lexical–semantic relations. In W. B. Croft, & C. J. van Rijsbergen (Eds.), *Proceedings of the 17th annual international ACM-SIGIR conference on research and development in information retrieval, Dublin, Ireland, July 3–6, 1994 (special issue of the SIGIR Forum)* (pp. 61–69). ACM/Springer-Verlag.

Wolfram, D., Spink, A., Janses, B. J., & Saracevic, T. (2001). Vox populi: the public searching of the web. *Journal of the American Society for Information Science and Technology, 52*(12), 1073–1074.

Xu, J., & Croft, W. B. (1996). Query expansion using local and global document analysis. In H.-P. Frei, D. K. Harman, P. Schäuble, & R. Wilkinson (Eds.), *Proceedings of the 19th annual international ACM SIGIR conference on research and development in information retrieval, Zurich, Switzerland, August 18–22, 1996 (special issue of the SIGIR Forum)* (pp. 4–11). ACM.

Xu, J., & Croft, W. B. (2000). Improving the effectiveness of information retrieval with local context analysis. *ACM Transactions on Information Systems, 18*(1), 79–112.

Xu, J., Fraser, A., & Weischedel, R. (2001). TREC 2001 Cross-lingual retrieval at BBN. In E. M. Voorhees, D. K. Harman (Eds.), *The tenth text retrieval conference (TREC 2001)* (pp. 68–77). NIST Special Publication 500-250.

Zazo, Á. F., Figuerola, C. G., Berrocal, J. L. A., & Gómez, R. (2002). *Recuperación de información utilizando el modelo vectorial. Participación en el taller CLEF-2001*. Technical Report DPTOIA-IT-2002-006, Departamento de Informática y Automática—Universidad de Salamanca. Available: http://tejo.usal.es/inftec/2002/DPTOIA-IT-2002-006.pdf.

ELSEVIER

Data & Knowledge Engineering 37 (2001) 25–45

DATA & KNOWLEDGE ENGINEERING

www.elsevier.com/locate/datak

Thesaurus construction through knowledge representation

Sean Bechhofer *, Carole Goble

Information Management Group, Computer Science Department, University of Manchester, Manchester, Oxford Road, Manchester M13 9PL, UK

Received 28 February 2000; received in revised form 9 May 2000; accepted 14 November 2000

Abstract

Semantic metadata describing subject content plays a vital role in supporting indexing and retrieval in Digital Libraries. Mechanisms used to deliver this metadata include keyword collections, thesauri and classifications. Constructing a large thesaurus, however, is a difficult process which can be facilitated through the application of knowledge representation techniques developed for managing and reasoning about concepts. We describe such a scheme – a Description Logic (DL) – and show through an example how a DL can play a part in the classification construction process, aiding in the production of coherent hierarchies and ensuring that the relationships represented in a thesaurus are sensible. © 2001 Elsevier Science B.V. All rights reserved.

Keywords: Description Logics; Thesauri; Subject-based classification

1. Introduction

Semantic metadata, information describing the content of documents, is a major subject of interest for Digital Libraries. Metadata extends to all manner of types and covers a wide range of areas including content of subject description; creation or bibliographic information and so on [1]. Our interest here is in subject classification and the use of content description in supporting searching and retrieval. This classification information is generally human created, but may also be elicited through analysis or summaries.

Current approaches to content metadata typically use mechanisms such as keywords. There are at least two problems with keywords, however, as there is no standard for their use, and consistent keyword descriptions are elusive. Controlled vocabularies and thesauri are an attempt to limit the terms and bring together index and query expressions.

1.1. Thesaurus-based retrieval

When used during retrieval and searching, thesauri are useful in bridging the gap that exists between the metadata provided by the indexer and the concepts presented by a searcher [2]. The

* Corresponding author. Tel.: +44-0161-275-6145.
E-mail address: sbechhofer@cs.man.ac.uk (S. Bechhofer).

0169-023X/01/$ - see front matter © 2001 Elsevier Science B.V. All rights reserved.
PII: S0169-023X(00)00052-5

controlled vocabulary limits the terms available and increases the possibility that the query will use appropriate terms. If the thesaurus has structure in the form of relationships such as broader or narrower terms (BT/NT), these may also help the searcher in navigation through the metadata and finding an appropriate query expression. If a query is too inclusive, then narrower terms may be substituted to refine the query. On the other hand, a query returning too few results can be broadened through the substitution of broader terms. Related terms (RT) may also be an aid in navigation and query construction. Search systems may help the user by including, say, narrower terms automatically (or at the request of the user).

Thesaurus construction is practised as an art – collections are generally crafted by hand (albeit according to some rules or guidelines) leading to problems with the consistency of their interpretation. If thesauri, and in particular, the *relationships* within thesauri are to be used for retrieval as suggested above, a consistent and well-understood interpretation or *semantics* is required for the relationships. This is particularly important if query expansion is to be automated.

As an example, consider the hierarchy shown in Fig. 1. This is a section of the ICONCLASS classification [3], and shows some of the terms occurring immediately below Door. ICONCLASS itself is not a thesaurus as it lacks some of the structure discussed above, but it does serve as an illustration of the difficulties of constructing classification hierarchies. Although there is one term which is truly a "kind-of" Door (Monumental door), the hierarchy includes many other relationships, including partonomy – Door-Knocker is a part of a door – and association – Closing the door is an action applied to a Door. This confusion of the role of the hierarchical relationship is common [4]. If the relationships are then used to expand or manipulate queries, unexpected results may ensue.

Svenonius considers the possibility of using "knowledge representations embodying classificatory structures" [5]. As Svenonius points out, this does not obviate the need for intellectual input, but the support offered by the knowledge representation can ease the burden of the construction process. In the same paper, the possibility of related terms being inferred is discussed, with the suggestion that "it may be that related-term relationships derived according to some rule might really be more productive in retrieval than those created through a subjective or scattershot approach". This is the approach we advocate and we present a mechanism which can assist in the production of related term links.

In this paper we describe a knowledge representation (KR) scheme known as a Description Logic (DL), and show that the use of such a scheme can aid in the construction of a coherent term collection. In particular, the DL can support a descriptive approach to modelling, resulting in

41A32 Door
- 41A322 Closing the door
- 41A323 Monumental door
- 41A324 Metalwork of a door
  - 41A3241 Door-knocker
- 41A325 Threshold
- 41A327 Doorkeeper, houseguard (inanimate)

Fig. 1. Fragment of ICONCLASS hierarchy.

clear hierarchical and cross-hierarchical relationships. We are particularly interested here in the *relationships* between terms as these can help support such activities as navigation, query expansion and similarity-based searching. We argue that the use of a compositional model, backed up by sound reasoning, can help in the construction of a coherent classification, and supply the required consistent semantics for links and relationships.

We do not intend to claim that DLs provide some panacea which will solve all the problems of indexing and searching, rather that the use of such a representation as one of a portfolio of retrieval services can be beneficial in particular circumstances. Methodologies for ontology construction are clearly also required, but the well-behaved classification and reasoning services provided by a DL are a useful addition to the toolbox of the ontology developer.

We begin with a statement of the problems we wish to address, and a brief overview of some existing solutions. This is followed by the description of DLs along with a discussion of how the characteristics of a DL can support the thesaurus construction activity. We then give a presentation of some examples of building a thesaurus using such an approach, and close with a discussion of related work and pointers to future research directions.

This work formed part of the Structured Terminology for Archives (STARCH) project [6], where we investigated how one might use and apply KR techniques to improve indexing and retrieval. Our examples use a terminology of costume which is being used in our prototype applications in collaboration with the City Art Gallery in Manchester.

2. Classifications and thesauri

We consider a *thesaurus* to be a collection of terms along with some structure or relationships between them as described by Aitchison and Gilchrist [7]. There are a number of relationships that might be represented here, including broader/narrower terms and associated or related terms. Standards [8] exist that set down the kinds of relationships that a thesaurus should represent.

Three issues pervade work in this area:

- Classification.
- Term composition or synthesis.
- Associative relationships.

2.1. Classification

A *classification* is a collection of terms divided or organised into subclasses, which may or may not be based on the kind-of relationship – Marcella and Newton [9] describe a classification as being a systematic arrangement of index entries in such a manner as to be useful for those seeking information. Such schemes (e.g., Dewey Decimal) have been used for many years in traditional library catalogues. A *hierarchical classification* is a collection of terms along with a relationship that represents a classification or kind-of hierarchy.

Note that these descriptions may overlap – many thesauri are also classifications. For example, The Art and Architecture Thesaurus, or AAT [10] is a thesaurus and classification, but is not really a hierarchical classification (although it is sometimes used as such). The BT/NT links are not pure classification as they represent other relationships – for example, the term people is a broader term of groups of people. The Dewey Decimal Classification System is a classification, while WordNet [11] can be considered a thesaurus.

The notion of hierarchical classification is, in our opinion, crucial if the vocabulary is to be used in retrieval, particularly if the retrieval process is to include query expansion or navigation.

2.2. Composition

In addition to classification, the idea of composition or synthesis of terms is instrumental in supporting the requirements of users [7]. There are a plethora of techniques to support composition.

2.2.1. Pre-coordinate systems

In a pre-coordinate system, terms are combined into a linear string which is then used as the index. Thus a search term has to combine the terms in the correct order. We can achieve broad searches by looking at the head terms of the index string, but this only works with one term. As the search becomes more precise, recall suffers as any differences in the order of the terms suppresses the search. We can put restrictions on the order that terms should be used (the so-called citation order) to help with this. One solution would be to record all permutations but this may make the index too large.

2.2.2. Post-coordinate systems

In a post-coordinate system, a document can have many index terms assigned to it – the terms are not combined, but remain independent. When searching, requests use combinations of terms which are matched against the index terms. Post-coordinate systems can use boolean operators such as AND, OR and NOT, word fragment searching or pattern-matching. *Links* are used to show which terms are being combined in the same documents – the combination is indicated by a grouping of the terms in the index. Links improve precision, but can affect recall. *Roles* are attached to terms and indicate the use or sense of the term. However roles are difficult to apply consistently (both in indexing and searching).

Factoring splits a term into its compound parts [7]. These are then either used as components of a pre-coordinate index or as elements in a post-coordinate index. Factoring can help to increase the recall performance of an index, but may effect precision. On the other hand, maintaining compound terms may help precision but leads to problems with maintenance of the thesaurus and index.

2.2.3. Faceted classification

Faceted classification, as described by Vickery [12], is based on the notion of concept coordination, where subject matter is represented by the coordination of two or more symbols. Thus a faceted classification is rooted firmly in the notion of combination. The terms are sorted into groups and hierarchically organised. Faceted schemes are primarily used in precoordinate systems. Using a faceted approach helps in several ways. The *Classificationist* (or modeller) starts with a large collection of terms that need to be organized. Many are compound, and require factoring as described above. The facets are an aid to this task, indicating ways of dismantling concepts. The *Classifier* or indexer is aided as the facets identify the structure of the compound terms and also identify the order that terms should be composed (which is important to improve recall in precoordinate systems). Finally, the *Searcher* is aided as the facets can help identify possible combinations to be presented as query expressions.

If the terms are to be used in a precoordinate system, we need to supply an order or sequence of facets which specifies how the terms are used in the index – in effect suggesting ways in which the composition should be applied.

2.3. Associative relationships

Along with classification and composition, a third device – *associative relationships* and RT links – is employed. Thesaurus standards [8] describe various situations where RT links should be present, and Lancaster [13] provides a discussion of the various categories of associative relationships. These including examples such as: an occupation and the person in that occupation, e.g., ACCOUNTANTS and ACCOUNTANCY; a thing or action and its counteragent, e.g., PESTS and PESTICIDES and an action and its product, e.g., ROADMAKING and ROADS. These RT links are then available for use in navigation and query expansion or construction.

2.4. Ontologies

A vocabulary, which can be considered to be a classification, terminology or thesaurus, should have an underlying *ontology*, defined by Guarino [14] as "an intensional semantic structure which encodes the implicit rules constraining the structure of a piece of reality". A *formal* ontology has some underlying logical structure which allows us to *reason* about the concepts in the ontology. It is this reasoning that we exploit in order to add coherency to the hierarchies which we construct.

2.5. Problems with thesauri

There are a range of problems caused by the fact that there is no *consistent* interpretation for relationships between terms and that the composition of terms effectively takes place outside the vocabulary. Devices such as links and roles or some specified order of composition attempt to impose some interpretation on this composition, but this is in an ad hoc way. The coordinated terms synthesised are independent of the classification – the composition and hierarchy construction are loosely coupled, if at all.

In many situations, classifications are naturally *graphs* rather than trees and the use of single inheritance can lead to problems as choices have to be made as to where terms should fit in a hierarchy. Maintenance of multiply inherited classifications, however, is hard to do without support.

Support for *incremental* change and development is required, particularly if the collection is being built bottom up, with the classification structure being derived piecemeal. The maintenance of collections or addition of extra terms may also require changes in the classification. The use of rigid asserted classification structures makes this process difficult.

Completeness is a problem with associative relationships. For example, in the AAT, there are RT links from VIOLINISTS to VIOLINS and GUITARISTS to GUITARS and so on, but each must be explicitly introduced. Ensuring that all instrumentalists are linked to their corresponding instrument is a difficult, tedious and expensive task. Again, such situations can lead to repetition in the hierarchies, in the sense that the musicians hierarchy and the instruments hierarchy will share a very similar structure.

We suggest that through the use of a representation which naturally supports composition of concepts and their classification within some unifying framework, we can supply a vocabulary with a consistent interpretation. Such a representation is supplied by a DL.

3. Description logics

DLs are a family of class-based knowledge representation languages stemming from KL-ONE (for an overview of the KL-ONE family of languages see [15]), allowing the construction and representation of conceptual models. A DL model is based on the notions of *concepts*, which represent classes of objects with similar characteristics, *individuals* which are instances of concepts, and *roles* which are relationships between individuals. Central to a DL are the notions of *subsumption* and *classification* – one concept is said to be subsumed by another when all of its instances are necessarily instances of the subsumer. Subsumption allows the construction of a classification hierarchy, with conceptual definitions being arranged from the general to the specific.

3.1. Classification and composition

A DL model is grounded on a collection of primitive concept definitions, along with assertions about the subsumption relationships between those primitive definitions. This is similar to a traditional asserted hierarchical classification structure. Where a DL differs from a thesaurus or classification scheme, however, is that the language provides a number of concept forming operators for *combining* primitive concepts and roles to form new concept definitions, along with a collection of reasoning services which allow us to make inferences about the interpretations of those *compositions*. In particular, if we construct a new composite, it can be *classified*, i.e., its position in the subsumption hierarchy is determined automatically. Furthermore, this reasoning is based on sound logical principles. The basic operators present in a core DL are shown in Table 1 (both the concrete and abstract forms are given based on those used by Baader et al. [16]). The semantics of the operators is also shown, grounded on collection of objects – the domain – and an interpretation function $(\cdot)^I$ which maps concept names to subsets of the domain and maps composite expressions according to the rules shown.

3.2. Reasoning services

A terminological knowledge base Σ contains a number of terminological axioms (see Fig. 2) which define relationships between primitive concepts and specify additional information about compositions. The axioms place conditions on the set of possible interpretations as shown in the table. Given a terminological knowledge base Σ, a DL provides a number of *reasoning services* – inference mechanisms which can deduce implicit knowledge from that given explicitly by the user.

Subsumption. Given two concept descriptions C and D, C is said to subsume D, $D \sqsubseteq C$, when it is necessarily the case that all the instances of D are instances of C.

Classification. Using subsumption, we can build a classification lattice of concept definitions. The classification is minimal with respect to subsumption relationships, thus if A subsumes B and B subsumes C there will be no direct link between A and C.

Satisfiability. Given a concept description, we can check that the description is satisfiable, i.e., that it is possible to find a model in which the description has a non-empty interpretation. For example, the concept description $(C \sqcap \neg C)$ is unsatisfiable, as we cannot have a thing which is both a C and not a C.

In addition to the terminological or T-Box reasoning services that are described above, a DL implementation can also support reasoning over individuals (known as A-Box reasoning). The

Operator	Concrete	Abstract	Semantics
Top	TOP	$\top$	Δ^I
Bottom	BOTTOM	$\bot$	$\emptyset$
Conjunction	(and C D)	$C \sqcap D$	$\{c \mid c \in C^I \cap D^I\}$
Disjunction	(or C D)	$C \sqcup D$	$\{c \mid c \in C^I \cup D^I\}$
Negation	(not C)	$\neg C$	$\{c \mid c \notin C^I\}$
Existential	(some R C)	$\exists R : C$	$\{c \mid R^I(c) \cap C^I \neq \emptyset\}$
Universal	(all R C)	$\forall R : C$	$\{c \mid R^I(c) \subseteq C^I\}$
Atmost	(atmost n R C)	$\leq nR : C$	$\{c \mid R^I(c) \cap C^I \leq n\}$
Atleast	(atleast n R C)	$\geq nR : C$	$\{c \mid R^I(c) \cap C^I \geq n\}$
Exact	(exact n R C)	$= nR : C$	$\{c \mid R^I(c) \cap C^I = n\}$
Inverse Role	(inv R)	R^{-1}	$(R^{-1})^I(c) = \{d \mid c \in R^I(d)\}$

Axiom	Concrete	Abstract	Semantics
Introduction	(defconcept CN)	CN	$CN^I \subseteq \Delta^I$
Introduction	(defrole RN)	RN	$RN^I : \Delta^I \longrightarrow \mathcal{P}(\Delta^I)$
Functional	(functional R)	func R	$\forall c \in \Delta^I, \lvert R^I(c) \rvert \leq 1$
Transitive	(transitive R)	trans R	$\forall c, d, e \in \Delta^I, d \in R^I(c) \wedge e \in R^I(d) \Rightarrow e \in R^I(c)$
Inclusion	(impliesrole R S)	$RN \sqsubseteq RS$	$RN^I \subseteq RS^I$
Inclusion	(implies C D)	$C \sqsubseteq D$	$C^I \subseteq D^I$
Equivalence	(equivalent C D)	$C \doteq D$	$C^I = D^I$

Fig. 2. DL syntax and semantics.

knowledge base will then have a number of assertional axioms which express facts about individuals in the domain. A *retrieval* function allows us to retrieve all the instances of a particular concept, *instance checking* enables us to check whether an individual is an instance of a particular concept, and *realization* determines the most specific concepts (with respect to subsumption) the individual is an instance of. In this paper, we concentrate on T-Box reasoning, as this is the functionality which we are exploiting when constructing a conceptual model.

To illustrate these ideas, consider an example model which has the primitive hierarchies shown in Fig. 3, along with two relations madeFrom and wornOn. We can now build up a collection of descriptions, e.g.,

1. (and Item (some madeFrom NaturalMaterial));
2. (and Item (some wornOn Arm));
3. (and Item (some madeFrom Wool));
4. (and Item (some madeFrom Silk) (some wornOn Leg)).

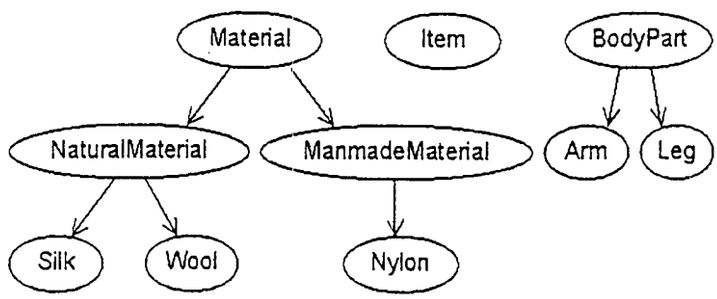


Fig. 3. Primitive hierarchies.

By examining the semantics of the descriptions as given in Fig. 2, we can deduce that any instance of (Item (some madeFrom Silk) and (some wornOn Leg) will also be an instance of (Item (some madeFrom NaturalMaterial)). This is the kind of inference that the subsumption checking process is able to make. The descriptions from the above example can then be classified as shown in Fig. 4. The important point to note here is that the classification is constructed automatically – the modeller does not have to explicitly position a composed concept such as (and Item (some madeFrom Silk) (some wornOn Leg)). In addition, the classification is dynamic – there is no need to introduce all required compositions before the classification is used. As new compositions are described, they will be placed in the appropriate place in the hierarchy. Multiple inheritance is supported in the classification, as in the example above, where (and Item (some madeFrom Silk) (some wornOn Leg)) is both a kind of (and Item (some madeFrom Silk)) and (and Item (some wornOn Leg)).

3.3. *The classification as index*

In the presence of A-Box reasoning, the conceptual hierarchy can be thought of as an index to the space of individuals. Retrieval takes into account the classification hierarchy, thus the classification encapsulates a hierarchy of query inclusions – if we make a query based on some high-level abstraction, all instances of subsumed concepts will be returned. An interesting point here is that we can easily form descriptions of higher level or abstract concepts, for example the notion of an Item made from NaturalMaterial, which subsumes both (Item made from Silk and Item made from Wool). The use of the DL allows us to build classification hierarchies which are consistent and coherent. *Multi-axial hierarchies* can easily be constructed (where each concept can have more than one parent in the classification), a task which is difficult to achieve by hand. Multi-axial classification can be especially useful if we want comprehensive coverage.

DLs have been proposed as a delivery mechanism for ontologies and vocabularies in the Tambis and GALEN projects [17,18], and have also been proposed as a mechanism for describing data sources within Digital Libraries [19]. For a more detailed description of DLs and their uses, see Borgida's overview [20,21].

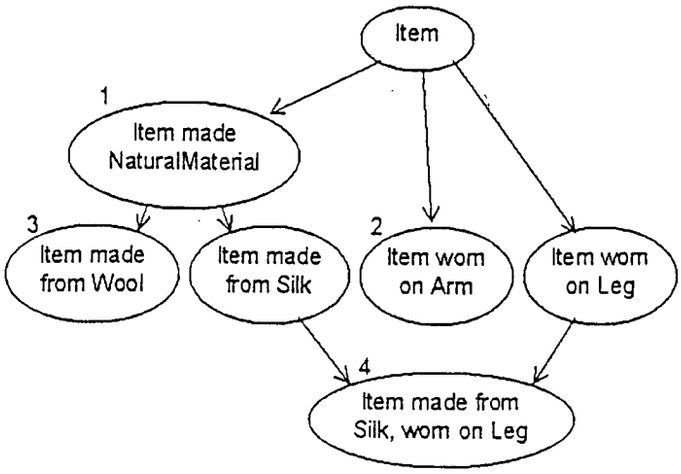


Fig. 4. Classification of descriptions.

3.4. The FaCT description logic

In the past, results concerning the intractability of reasoning within DLs have been used to dismiss their use in real world applications. Although DL languages are generally known to be intractable in the worst case, there has been much interest of late in the DL community in implementations of optimised reasoning engines which can deliver realistic performance for practical applications. One such logic is FaCT, a DL developed at the University of Manchester [22]. FaCT uses a tableaux-based reasoner along with sophisticated optimisation techniques to provide sound and *complete* reasoning for an expressive language. Work on FaCT is continuing, and has recently resulted in the addition of qualified number restrictions to the language [23], and a prototype A-Box implementation [24]. The current implementation of FaCT provides the expressivity of the *SHIQ* language [23] which includes all the operators shown in Fig. 2.

In our intended architecture, the DL is seen as a terminological resource, as discussed by the authors in [25]. An implementation of FaCT with a CORBA-based wrapper [26] is available for download from the University of Manchester Department of Computer Science [27]. FaCT is also being used as the representation language in a number of diverse research projects and efforts including Tambis [28], a system providing access to multiple information sources and DWQ [29] where the DL is used in schema integration and verification. FaCT provides the underlying reasoning services which form the basis of the Ontology Inference Layer OIL [30], further discussed in Section 5.

4. Modelling a thesaurus

The International Committee for the Museums and Collections of Costume (ICOM) produce a vocabulary of basic terms which is used for the description of pieces of costume [31]. The vocabulary is broken down into three main subdivisions, men's garments, women's garments and infant's garments, with each main division (i.e., men/women/infants) being further broken down into main garments, outerwear, underwear, etc. As it stands, there is much duplication in the vocabulary, and the repeated subdivision lends itself to a compositional treatment as discussed here. The AAT costume hierarchy is similar, with divisions according to form and function.

We have reworked and remodelled the ICOM vocabulary as a DL model. In addition we have incorporated some terms from the AAT and a number of keyword descriptors which have been used in the Platt Hall Gallery of Costume Collection, part of Manchester City Art Galleries. The costume gallery is a source of material for the prototype implementation of a system using DL-based models in retrieval being investigated as part of the STARCH project [6].

The vocabulary was built in a bottom up manner, using a collection of primitive relationships and concepts to define the concepts in a compositional manner. Thus far, approximately 300 concepts have been defined in the vocabulary.

Representing the vocabulary as a DL model requires the use of the DL reasoning engine. It may not always be the case that the reasoner is the appropriate method for delivery of the terms. In many cases, a traditional static thesaurus may well be of more benefit. In this case, we can "export" a static representation of the thesaurus, using the relationships and reasoning services of the DL to decide on appropriate thesaurus links. This approach towards vocabulary construction has been used in the GALEN-IN-USE project [32], where a model represented in the DL GRAIL

[33] is used to help create coding schemes representing clinical terminology. Although the final thesaurus is no longer represented using the DL, the use of the DL has helped in the construction of the term collection, ensuring that the relationships such as BT/NT are coherent.

4.1. Classification and composition

To illustrate the use of composition, a Corset is an item which is worn both above and below the waist and has the purpose of support, a Shirt is a main garment which is worn above the waist and a Bracelet is a decorative item worn on the arm. Each of these can thus be defined in terms of the more primitive notions, i.e.,

```
Corset ⊑ (and Item (some purpose Support)
          WornAboveWaist WornBelowWaist)
Shirt ⊑ (and MainGarment WornAboveWaist))
Bracelet ⊑ (and Item (some purpose Decoration)
    (some wornOn Arm))
```

In addition, more abstract notions can also be defined. For example:

```
SupportGarment ⊑ (and Item (some purpose Support))
```

In this way, the classifier can take care of the subsumption (kind-of) relationship between the notion of a SupportGarment and a Corset. The idea of SupportGarment need not be introduced before Corset, but can be defined later – this is useful when the model is being constructed as we can work in a bottom up fashion, incrementally introducing aspects of the hierarchical structure, and allow the reasoner to deal with the burden of maintaining consistency within the hierarchies. There is no need for the modeller to reorganize the classification hierarchies as they change, as this is dealt with by the classifier.

4.2. Expressiveness and definition through description

The concept forming operators in FaCT include disjunction ("or") allowing us to model combinations of concepts in a natural fashion. As an example, consider men's suits. ICOM gives a number of different descriptions of suits, characterised by the pieces that they are composed of. For example, a suit could be made up from three pieces – coat waistcoat and trousers – or could be simply a coat and waistcoat.

We can use the or operator in conjunction with numerical restrictions to provide the definitions of the concepts as shown in Fig. 5. To explain the descriptions, we are saying that Suit1 is composed of three things which are a Coat, Trousers and Waistcoat, while Suit2 is a Coat and Trousers and so on. A Suit, in general, is any of the given combinations. The disjunction operator allows us to represent this precisely – in thesauri, it is not always clear whether a combination of terms represents conjunction or disjunction. This arbitrary semantics may lead to confusion and misinterpretation of terms. Through the use of the DL we can see *exactly* what is meant by each definition. In the ICOM classification, the fact that Suit1 is composed of a Coat, Trousers and Waistcoat is dealt with through the use of a scope note or description. In this case, through the description, we are *explicitly* representing information which is *implicitly* provided via the scope note.

```
Suit1 ≐ (and Combination (exact 3 composed_of Item)
        (some composed_of Coat)
        (some composed_of Trousers)
        (some composed_of Waistcoat))

Suit2 ≐ (and Combination (exact 2 composed_of Item)
        (some composed_of Coat)
        (some composed_of Trousers))

Suit3 ≐ (and Combination (exact 2 composed_of Item)
        (some composed_of Coat)
        (some composed_of Waistcoat))

Suit4 ≐ (and Combination
        (exact 2 composed_of Item)
        (some composed_of Jacket)
        (some composed_of Trousers))

Suit ≐ (or Suit1 Suit2 Suit3 Suit4)
```

Fig. 5. Definitions of suits.

The classifier will now take care of the subsumption (BT/NT) relationships between the concepts, inferring that Suit1 is a kind of Suit2. More interestingly, however, is the fact that Suit2, for example is now a kind of (some composed_of Trousers), so if we wish to introduce this concept to represent the notion of a Combination which includes Trousers, Suit2 will be subsumed by this concept.

The AAT contains the term Tuxedo, which is described (through the scope note) as being a combination of a Dinner_Jacket and Trousers, where a Dinner_Jacket is a kind of Coat (but not, surprisingly a Jacket). If we add the definitions

```
Dinner_Jacket ⊑ Coat
Tuxedo ⊑ (and Combination
        (exact 2 composed_of Item)
        (some composed_of Dinner_Jacket)
        (some composed_of Trousers))
```

to the knowledge base, we find that a Tuxedo is now classified as a kind of Suit2. The resulting hierarchy is as shown in Fig. 6.

This example illustrates how the relationship between constituent parts of a composition can influence the relationship between that composition and other descriptions – here the relationship between Dinner_Jacket and Coat induces a relationship between Tuxedo and Suit2. If we were to change our view of Dinner_Jacket and decide that it was, in fact a kind of Jacket rather than a Coat, we need only change the axiom

```
Dinner_Jacket ⊑ Coat
```

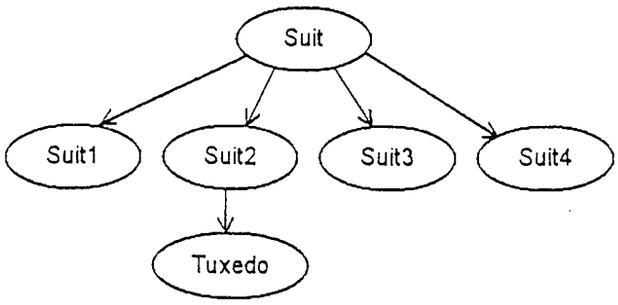


Fig. 6. Hierarchy of suits.

to

Dinner_Jacket $\sqsubseteq$ Jacket

The classifier would then take care of the positioning of Tuxedo on the hierarchy (in this case it would then be a kind of Suit4 rather than Suit2). The reorganization of the hierarchies is being controlled and governed by the classifier – this illustrates how the classifier can help the management of change and the maintenance of a *consistent* hierarchy.

4.3. Methodologies

Of course, adopting a DL as the underlying representation does not instantly solve all problems of ontology construction! Methodologies which guide the modeller in deciding exactly how to break down the world (for example do we introduce a primitive notion of WornAboveWaist; WornBelowWaist or a role worn along with the appropriate fillers AboveWaist and BelowWaist?) are required. The use of a DL does not abdicate the modeller from the responsibility of choosing a representation which is appropriate.

The use of the classifier and reasoner, however, can smooth the process, allowing the modeller to explicitly *describe* the world and have some of the resulting structure inferred through those descriptions.

4.4. Multiple classification

A problem with a simple thesaurus is that of multiple classification. In many cases, allowing multiple views of a concept would be useful. For example, in the AAT, the term Raincoat appears under <costume by form> :outerwear:overcoats:raincoats. One could argue that a rain coat should also be considered an item of protective wear. Within the scope note for protective wear, there is a comment to the effect that for garments worn for protection against weather, the term outerwear or its narrower terms should be used. Such a mechanism is less likely to be of use to computer-based search. Similarly, the term running shoes appears under <shoes by function> which is itself under <footwear by form>. One might also expect running shoes to appear somewhere in the hierarchy under <footwear by function>, but this is not the case.

It would be more fruitful to allow the multiple classification of Raincoat as both an item of outerwear and a protective garment or to allow running shoes to appear in multiple positions

within the hierarchy. However, without the classification support offered by some KR scheme, such multi-axial hierarchies are very hard to maintain.

4.5. Related terms

In the case of the ICOM model, the subsumption relationship corresponds to BT/NT links. For RT links, we can examine the description applied to a particular definition. For example, a Helmet is defined as an item which is worn on the Head and is worn for the purpose of Safety:

```
Helmet ⊑ (and Item (some purpose Safety)
  (some wornOn Head))
```

Thus Helmet will be classified as both a SafetyItem and an article of Headwear. In addition, the classifier can deduce that Head and Safety are related to Helmet as they are used in its definition.

We can now introduce the ideas of activities or occasions, and describe pieces of clothing associated with those activities or occasions. For example, if we introduce an idea of Motorcycling as an Activity, we can now define:

```
MotorcycleHelmet ≐ (and Helmet (some worn_during Motorcycling))
MotorcycleGlove ≐ (and Glove (some worn_during Motorcycling))
```

In this way, Motorcycling is deduced to be an RT of both MotorcycleHelmet and MotorcycleGlove. If desired, the related terms devised in this way could be inherited along the BT/NT links. In the example above, the term Glove may well have been defined as being a kind of item which is worn on the Hand, leading to an induced relation between Hand and Glove. This relationship could then be inherited by MotorcycleGlove, leading to an RT link between MotorcycleGlove and Hand as well as MotorcycleGlove and Motorcycling.

In the example of Suits above, we can infer that Jacket is related to Suit2, through the use of Jacket in the description of Suit2. The possible benefit of related terms being inferred in this way is raised by Svenonius [5].

As discussed by Lancaster [13] and referred to earlier, there are a number of different categories of associative relationships which may occur. Through the explicit use of the relationships and roles of the underlying knowledge representation, the particular associative relationship can be made clearer. In the Helmet example above, we can see that Safety is a related term because it is the purpose for which the Helmet is worn.

4.6. Collapsing semantics

As discussed in Section 2.3, there may be many different interpretations of the RT links which are being explicitly represented within the DL. Examples include partitive relationships (a Suit is composedOf a Jacket), or usage relationships (a Glove is wornOn the Hand). When translating to the thesaurus, however, the semantics of the relationships are being collapsed down to a single notion of *related to* (the RT link). We must be careful with this translation.

This is an inherent problem with the expressivity of the thesaurus structure, however, as it only provides a *single* notion of RT. Regardless of the methodology or approach used to produce the relationships, the paucity of relationships within the thesaurus structure leads to this difficulty. Although notes and guidance as to the use of the thesaurus can be given – for example through the

scope notes for the terms – if the thesaurus is being used within an retrieval system, these notes may not be accessible or interpretable by the system. This suggests to the authors that a richer representation is needed in order to capture more precisely the meanings of the different relationships and to aid their use within a system. The approach discussed within this paper is but the first step towards this goal.

As discussed above, a principled approach is also required here as we may not expect every relationship to induce a related term and this is clearly an area where a methodology is key. If a role hierarchy is present, there may be high-level roles (for example some high-level partonomic relationships) which we do not wish to use in this way. Again, though, the underlying DL representation and explicit representation of the relationships between the roles provides us with a "conceptual coatrack" on which we can hang our requirements.

4.7. Bottom up construction

The exercise here is a form of "bottom up" thesaurus construction similar to that used in building a faceted classification [12]. Starting with a collection of terms provided by ICOM, the hierarchical structure of the thesaurus has been constructed. Through the use of the classification functionality provided by the DL, this structure is in the main *induced* through the description and definition of the composed concepts.

The vocabulary produced using this technique can be viewed on the Web via Java applets which allow access to the static vocabulary produced as described above [34]. Fig. 7 shows a screen shot of the simple thesaurus viewer.

4.8. Pre- and post-coordination

The approach described here makes strong use of the idea of composition, along with automatic classification to govern the construction of hierarchies. This corresponds to ideas of pre-coordination as terms are combined to produce a single indexing term. There are elements of a

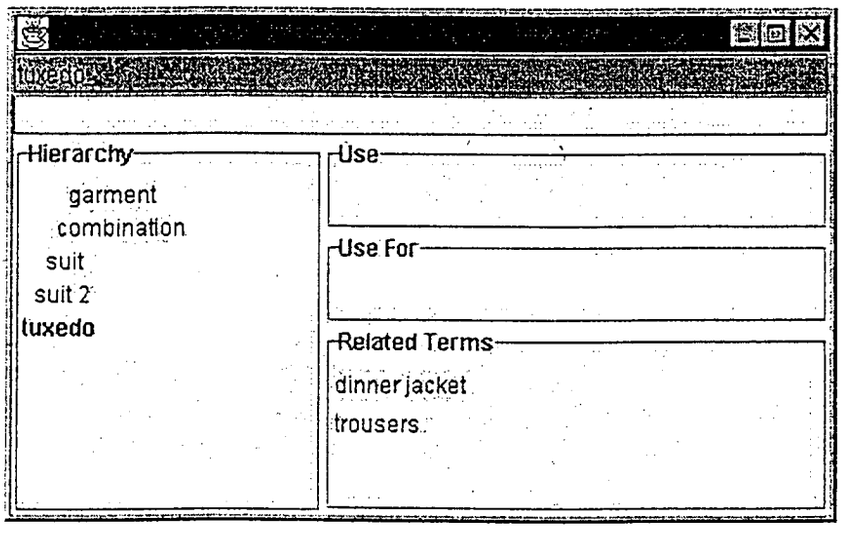


Fig. 7. Thesaurus term viewer.

post-coordinate approach here, however, as the order of the composition is unimportant (operators such as conjunction are commutative in a DL). In addition, we can use the hierarchy to allow the phrasing of abstract or general concepts.

In the DL model, combinations are within the context of a role, thus providing more meaning for the juxtaposition of the concepts. Similarly, the richer concept combination operations of the DL allow better representations of *how* and *why* concepts are being combined in a compound expression, providing a more principled mechanism that that given by, e.g. links and roles.

The compositional DL model can be considered as being broken down into individual facets, corresponding to top level abstract concepts, with concepts combined to provide compositions. When building DL models, much thought goes into the identification and construction of suitable high-level divisions of the model [18]. Systems of constraints, such as sanctions [6] can help to identify how the concepts should be combined.

5. Related work

Much recent metadata work seems to be concerned with four things that facilitate data sharing and interoperability:

1. a standard set of tags or elements that describe digital or physical objects, e.g. the Dublin Core discussed by Cathro [35] or Library of Congress Machine-Readable Cataloging [36];
2. rules for making and amending tags (for example XML);
3. methods for identifying standard vocabularies which will provide the terms used in the elements describing objects;
4. a container architecture [37].

Such work is complementary to that described here, where our thrust is to provide a framework for the consistent description of content – the terms would then be used as values in a particular field of the catalogue record.

Glamorgan's Semantic Hypermedia Architecture [38] uses a binary relational semantic data model as a semantic index space to a hypermedia system. The model is based on a static classification. However, their approach does not support terminological reasoning or dynamic classification – any composition of concepts will have to be along the lines of pre- or post-coordination in the traditional thesaurus sense.

In the context of the Glamorgan work, Alani [39] discusses problems caused by the lack of structure in the RT relationship when using traditional thesauri during retrieval. In the construction of a thesaurus such as the AAT, editorial information may be recorded as to why certain terms have been marked as related. Alani calls for a richer semantics and explicit representations of the RT sub-relationships – precisely the kind of support which an expressive DL can provide.

The Ontobroker project [40] proposes the use of a formal ontology to represent metadata. However, the representation (Frame-Logic) is static and does not offer the kinds of iterative classification provided by a DL.

Meghini [41] adopts a DL for information retrieval. However, their approach differs from ours in that they intend modelling both form (syntactic metadata) and content (semantic metadata) in one framework. This requires extensions to the DL formalism to deal with particular concrete

domains. In contrast, we advocate the "simple" application of a DL to assist in building coherent classifications of subject content.

An approach to using a DL to represent metadata is described by Weinstein [42]. Surprisingly, Weinstein claims not to use the classification reasoning services of the DL (described below) to generate the knowledge base, and further states "...nor is [automatic classification] likely to be important to support end-user queries". It is our view that this reasoning is the greatest asset of DL representations and is vital in supporting coherent models which will bridge the gap between indexer and searcher.

In [43], Jannink and Wiederhold describe an approach to thesaurus construction which relies on the automated analysis of an on-line dictionary. A term is considered related when it appears in the definition. An algorithm is then used to rank the relationships and uncover the *strongest* relationships. The method does not supply any information about the type of the relationship, simply a measure of its importance. This has some similarities with our approach, as we suggest using the constituent parts of the description of terms as pointers to related terms. The explicit structure in the DL descriptions provides more information about the types of relationship (with of course a corresponding increase in effort required to build the thesaurus). As Jannink and Wiederhold say, their approach is not intended as a replacement for handcrafted thesauri, but could be considered an adjunct to such devices.

Frank et al. [44] discuss the construction of a knowledge base using the CIA World Fact Book as a source. This is partly an automated process, using the structure of the original source to aid the knowledge extraction, but also involves the manual organization of terms into hierarchies. The authors discuss the alternatives of using several orthogonal (multiple) classifications or describing the terms using attributes. This can be supported by knowledge representation allowing the introduction of axioms which equate descriptions. The authors also briefly mention that attempts were made to capture relationships between terms during the modelling process – for example hydropower potential as a natural resource is used by the energy industry– but do not describe in detail how this was done.

Delivery of ontologies or vocabularies through services has been a focus of interest in several communities, with specifications of services such as the OKBC [45] and the OMG LQS [46]. The architecture of a Terminology Server is discussed in [25].

A recent development in languages for semantic metadata is the Ontology Inference Layer OIL [30,47], a proposal for a metadata language and standard for the Web. OIL draws on frame modelling primitives, but backs these up with a well defined semantics defined in terms of an expressive DL (*SHIQ* with the addition of concrete domains). This allows the provision of reasoning support which can aid the modeller in constructing coherent models. This philosophy of employing DL reasoning services underneath a "friendly face" ties in closely with the approach advocated here.

In [48], Welty discusses the use of DLs in catalogues and analyses a number of approaches. In particular, problems are encountered due to attempts to represent the catalogued items as individuals belonging to particular classes and properties such as "aboutness" as relationships between those individuals. In contrast to this, our suggestion here is that rather than using the DL as a primary delivery mechanism for the ontology or catalogue, the reasoning services can be used to support the construction of the model. As cited above, this was the successful approach adopted in the GALEN-IN-USE project [32].

6. Discussion

We have advocated the use of knowledge representation to support the construction of structured controlled vocabularies.

6.1. Benefits

Controlled vocabularies such as thesauri and classification schemes have proved invaluable tools to aid the processes of cataloguing and retrieval. If the vocabulary has some structure this can add to the power of the index, aiding the searcher in his or her task. However, in many situations this structure is ad hoc – without a consistent interpretation of the relationships, searching may well provide unpredictable results.

DLs provide a KR scheme that can deliver vocabularies in a principled manner, ensuring a consistent interpretation of the concepts. The support for composition and (automatic) classification make DLs an ideal candidate representation for a controlled vocabulary.

The dynamic classification allows the hierarchy to be built in an incremental fashion. Change and evolution can be supported, without the requirement for wholesale reorganisation of hierarchies as the classification is induced based on the descriptions of composite concepts. Multi-axial hierarchies – a useful device when building conceptual models – are easily supported by the DL classifier.

As discussed by Bates [2], being able to refine a query through navigation allows the searcher to explore and "feel their way", learning more about the model as he or she goes. This has also been identified by Bullock and Goble [49], where ethnographic studies isolate a need for spontaneity in information systems – the user's need evolves as more information is revealed. The use of a coherent and rigorously organised classification as the backbone to the organisation of the model and the relationships within it facilitates this navigation and browsing.

In addition to supporting the construction of classifications, the DL reasoner can also assist in identifying incoherent concepts through the use of satisfiability testing. This allows the modeller to identify contradictory concept descriptions – the identification of such inconsistencies is again a non-trivial task, particularly when expressive languages are in use.

The DL reasoner may also be used to control interaction with the underlying concept definitions, supporting the definition of forms-based interfaces [50,51]. Within the context of thesauri, this may help in providing alternative mechanisms for navigating the term collections other than through traversal of hierarchies and selection from picking lists.

6.2. Limitations

Simply using a DL does not solve all our problems because the use of knowledge representations does not remove the need for intellectual input. The conceptual models must still be constructed – experience in the GALEN [18] and TAMBIS [17] projects suggests that model building is a difficult process requiring modellers skilled in both the domain area and the representation language. Support tools are vital to this process.

It is also unreasonable to expect modellers to work directly with the DL formalism. The GALEN In Use project has had some success with the use of an intermediate representation [32] which helps to insulate modellers from the underlying syntax of the DL representation. However, the power of the DL is still being employed through the use of the classification and consistency

checking reasoning services. A similar approach is used in the Data Warehousing Quality project [29], where a tool translates an ER schema into a DL model in order to allow the definition and consistency checking of inter-schema constraints.

Investigations are required into the possibilities of reformulating existing classification schemes (for example AAT) into a compositional model, and the tools required for such a task.

In the example presented in this paper, the subject area deals with objects which are relatively easily broken down or faceted into purpose, material, form and so on. Such a domain is well suited to a compositional approach, and shares properties with the subject areas in which DLs have been successfully applied (medical terminologies, software engineering, configuration management). In other domains where such a decomposition is not so clear – for example the ASIS Thesaurus of Information Science [52], the DL modelling style may be harder to apply.

The computational complexity of DL reasoning is sometimes cited as a drawback. However, empirical evidence suggests that worst-case complexities rarely arise in *real* knowledge bases [22]. The production of efficient reasoning algorithms for DL representations is a topic of current research [53]. If we extend the use of the DL with A-box functionality, there are some limitations in the kinds of query which are supported.

However, our claim is not that DLs provide some panacea which will solve all the problems of indexing and searching, rather that the use of such a representation as one of a portfolio of retrieval services can be beneficial in particular circumstances.

6.3. Knowledge-based retrieval

A DL can aid in the process of building a controlled vocabulary. This is of course only exploiting some of the power of the DL. In the above examples, the DL is used to produce a static collection of terms (with an associated hierarchy). Further benefits are obtained if we take the approach a step further, dispense with the static thesaurus, and use the DL as an access mechanism [6]. This then allows us to exploit the dynamic nature of the DL reasoning engine, with query expressions formulated and classified on the fly.

As discussed earlier, the DL can support arbitrary expressions, including abstractions such as the notion of a combination including trousers. If we use the DL directly, these abstract expressions can be used as queries, and do not need to be inserted in the "thesaurus" pre-hoc.

Subject-based navigation can play an important part in conceptual hypermedia systems such as Nanard and Nanard's MacWeb [54], where documents are indexed with conceptual terms with relationships between terms inducing links between documents. Ideas of navigation based retrieval are further explored in [6].

Bates [2] suggests that users can produce more powerful searches if initial topics or terms yield further possibilities for related subjects or classifications – such functionality may be supported by a DL model through the use of constraints which described *how* compositions could be formed. In [50] and [55] the authors describe a forms-based model-driven user interface which allows a user to incrementally specify a query in this manner, with the forms proving some direction for the user as required, facilitating navigation through the index space [6]. In this way, the user no longer has to interact with the raw classification used in the indexing (although there may be certain situations where it is beneficial to expose the classification explicitly). The naïve user who knows nothing of the classification used can be helped in their navigation and query construction – the information system plays more the part of an intermediary rather than a simple query answering machine.

With the addition of an A-Box reasoner, the DL can form a powerful and flexible query engine with the DL model providing a *space* around which a user can navigate, combining the activities of browsing and query as discussed in [6]. The provision of a true A-box is not, however, mandatory for retrieval (and can introduce problems as highlighted in [48]). Alternatively, the retrieval function can be replaced by a selection or lookup service – the approach adopted in the TAMBIS project [17].

Acknowledgements

This work was supported by EPSRC grants GR/L71216 and GR/M75426. The authors would like to thank Richard Giordano (Center for Innovation in Product Development, MIT) for his guidance and advice, and the reviewers of the paper for their comments.

References

[1] C.C. Marshall, Making Metadata: a study of metadata creation for a mixed physical-digital collection. In: ACM Conference on Digital Libraries: DL98, Pittsburgh, PA, ACM Press, New York, 1998, pp. 162–171.
[2] M. Bates, Indexing and access for digital libraries and the internet: human database and domain factors, J. Am. Soc. Inform. Sci. 49 (13) (1998) 1186–1205.
[3] H. van de Waal, ICONCLASS: An Iconographic Classification System. Koninklijke Nederlandse Akademie van Wetenschappen, 1973–1985.
[4] R. Brachman, What is-a is and isn't: an analysis of taxonomic links in semantic networks, IEEE Comput. 16 (10) (1983) 30–36.
[5] E. Svenonius, Unanswered questions in the design of controlled vocabularies, J. Am. Soc. Inform. Sci. 37 (5) (1986) 331–340.
[6] S. Bechhofer, C. Goble, Classification based navigation for picture archives, in: IFIP WG2.6 Conference on Data Semantics, DS8, Rotorua, New Zealand, Kluwer Academic Publishers, Dordretch, 1999, pp. 291–310.
[7] J. Aitchison, A. Gilchrist, Thesaurus Construction – A Practical Manual, second ed., Aslib, London, 1987.
[8] International Organization for Standardization. ISO 2788: Guidelines for the Establishment and Development of Monolingual Thesauri, second ed., ISO, Geneva, 1986.
[9] R. Marcella, R. Newton, A New Manual of Classification, Gwoer Publishing, Aldershot, England, 1994.
[10] T. Peterson, Introduction to the Art and Architecture Thesaurus, second ed., Oxford University Press, UK, 1994.
[11] C. Fellbaum, WordNet: An Electronic Lexical Database, MIT Press, London, 1998.
[12] B.C. Vickery, Faceted Classification Schemes, Rutgers Series on Systems for the Intellectual Organization on Information, Rutgers State University, New Brunswick, NJ, 1966.
[13] F.W. Lancaster, Vocabulary Control for Information Retrieval, Information Resources Press, Arlington, Virginia, 1986.
[14] G.N. Formal Ontology and Information Systems. in: N. Guarino, (Ed.), Formal Ontology in Information Systems, Proceedings of the First International Conference, Trento, Italy, 6–8 June, IOS Press, 1998.
[15] W.A. Woods, J.G. Schmolze, The KL-one family, Comput. Math. Appl. 23 (2–5) (1992) 133–177.
[16] F. Baader, H.-J. Bürckert, J. Heinsohn, B. Hollunder, J. Müleer, B. Nebel, W. Nutt, H.-J. Profitlich, Terminological Knowledge Representation: A Proposal for a Terminological Logic, Technical Memo TM-90-04, Deutches Forschungszentrum für Künstliche Intelligenz (DFKI), 1991.
[17] P.G. Baker, C.A. Goble, S. Bechhofer, N.W. Paton, R. Stevens, A. Brass, An ontology for bioinformatics applications. Bioinformatics, Bioinformatics 15 (6) (1999) 510–520.
[18] A. Rector, J. Rogers, P. Pole, The GALEN high level ontology, in: Medical Informatics Europe '96, Copenhagen, 1996 pp. 174–178.
[19] Y. Velegrakis, V. Christophides, P. Constantopoulos, Declarative Specification of Z39.50 Wrappers Using Description Logics. in: European Conference on Digital Libraries, ECDL, Paris, Springer, New York, 1999.
[20] A. Borgida, Description logics in data management, IEEE Trans. Knowledge Data Eng. 7 (5) (1995) 671–782.
[21] C. Goble, What have the Romans (and Germans) ever done for us? Invited Talk. in: DL2000, International Workshop on Description Logics, Aachen, 2000.
[22] I. Horrocks, Using an expressive description logic: FaCT or fiction?, in: A.G. Cohn, L.K. Schubert, S.C. Shapiro (eds.), Principles of Knowledge Representation and Reasoning: Proceedings of the Sixth International Conference (KR'98), Morgan Kaufmann, Los Alitos, CA, 1998, pp. 636–647.

[23] I. Horrocks, U. Sattler, S. Tobies, Practical reasoning for expressive description logics, in: Proceedings of the Sixth International Conference on Logic for Programming and Automated Reasoning (LPAR'99), Springer, Berlin, 1999, pp. 161–180.
[24] S. Tessaris, G. Gough, ABox reasoning with transitive roles and axioms, in: Proceeding of the 1999 International Workshop on Description Logics (DL'99), 1999.
[25] S.K. Bechhofer, C.A. Goble, A.L. Rector, W.D. Solomon, W.A. Nowlan, Terminologies and Terminology Servers for Information Environments, in: Eighth International Workshop on Software Technology and Engineering Practice – STEP97, London, IEEE Computer Society, 1997, pp. 484 – 497.
[26] S. Bechhofer, I. Horrocks, P.F. Patel-Schneider, S. Tessaris, A proposal for a description logic interface, in: DL99, International Workshop on Description Logics, Linköping, Sweden, 1999, pp. 33–36.
[27] Camelot Project Home Page. http://www.cs.man.ac.uk/~horrocks/Camelot/.
[28] P.G. Baker, A. Brass, S. Bechhofer, C. Goble, N. Paton, R. Stevens, TAMBIS: transparent access to multiple bioinformatics information sources. An overview, in: Intelligent Systems for Molecular Biology ISMB98, Montreal, Canada, 1998.
[29] D. Calvanese, G.D. Giacomo, M. Lenzerini, D. Nardi, R. Rosati, Description logic framework for information integration, in: Proceedings of the Sixth International Conference on the Principles of Knowledge Representation and Reasoning (KR'98), 1998, pp. 2–13.
[30] D. Fensel, I. Horrocks, F. van Harmelen, S. Decker, M. Klein, OIL in a nutshell, in: Proceedings of the 12th International Conference on Knowledge Engineering and Knowledge Management, Juan-les-Pins, France, 2000.
[31] ICOM Costume Classification. http://www.open.gov.uk/mdocassn/costume/vbt00e.htm.
[32] J.E. Rogers, W.D. Solomon, A.L. Rector, P.M. Pole, E. Zanstra, P. and van der Haring, Rubrics to dissections to GRAIL to classifications, in: Medical Informatics Europe '97, Thessalonika, 1997, pp. 241–245.
[33] A.L. Rector, S.K. Bechhofer, C.A. Goble, I. Horrocks, W.A. Nowlan, W.D. Solomon, The GRAIL concept modelling language for medical terminology, Artificial Intell. Med. 9 (1997) 137–191.
[34] University of Manchester Information Management Group Web Site. http://img.cs.man.ac.uk/.
[35] W. Cathro, Metadata: an overview, in: Standards Australia Seminar, 1997.
[36] Library of Congress Machine-Readable Cataloging. http://lcweb.loc.gov/marc/marc.html.
[37] C. Lagoze, C.A. Lynch, R.D. Jr. The Warwick framework: a container architecture for aggregating sets of metadata, technical Report TR96-1593, Cornell University, Computer Science, 1996.
[38] D. Cunliffe, C. Taylor, D. Tudhope, Query-based navigation in semantically indexed hypermedia, in: Hypertext '97, Southampton, 1997.
[39] H. Alani, C. Jones, D. Tudhope, Associative and spatial relationships in thesaurus-based retrieval, in: J. Borbinha, T. Baker (Eds.), Proceedings of the Fourth European Conference on Research and Advanced Technology for Digital Libraries (ECDL 2000), vol. 1923 of LNCS, Paris, Springer, Berlin, September 2000, pp. 45–48.
[40] S. Decker, M. Erdmann, D. Fensel, R. Struder, Ontology based access to distributed and semi-structured information, in: IFIP WG2.6 Conference on Data Semantics, DS8, Rotorua, New Zealand, Kluwer Academic Publishers, Dordretch, 1999, pp. 351–369.
[41] C. Meghini, F. Sebastiani, U. Straccia, The terminological image retrieval model, in: Proceedings of ICIAP-97, Ninth International Conference on Image Analysis and Processing, Firenze, Italy, Springer, Berlin, 1997, pp. 156–163.
[42] P.C. Weinstein, Ontology-based metadata: transforming the MARC legacy, in: ACM Conference on Digital Libraries: DL98, Pittsburgh, PA, ACM, New York, 1998, pp. 254–263.
[43] J. Jannink, G. Wiederhold, Thesaurus entry extraction from an on-line dictionary, in: Proceedings of Fusion '99, Sunnyvale, CA, July 1999.
[44] G. Frank, A. Farquhar, R. Fikes, Building a large knowledge base from a structured source: The CIA World Fact Book. IEEE Intell. Sys. 14 (1).
[45] V.K. Chaudhri, A. Farquhar, R. Fikes, P.D. Karp, J.P. Rice, Open knowledge base connectivity 2.0.3., 1998.
[46] Object Management Group, Lexicon Query Service RFP Response, OMG: Working Document, 1998.
[47] C. Goble, R. Stevens, I. Horrocks, OIL: a slick way to represent knowledge for bioinformatics and the web, in: EML Workshop on Management and Integration of Biochemical Data, 2000.
[48] C.A. Welty, J. Jenkins, Formal ontology for subject, Data Knowledge Eng. 31 (1999) 155–181.
[49] J. Bullock, C. Goble, TourisT: the application of a description logic based semantic hypermedia system for tourism, in: Ninth ACM Hypertext Conference, Pittsburgh, PA, 1998.
[50] S. Bechhofer, C. Goble, Using a description logic to drive query interfaces, in: DL97, International Workshop on Description Logics, Gif sur Yvette, 1997.
[51] S. Bechhofer, I. Horrocks, Driving user interfaces from FaCT, in: DL2000, International Workshop on Description Logics, Aachen, 2000.
[52] ASIS Thesaurus of Information Science. http://www.asis.org/Publications/Thesaurus/tnhome.htm.
[53] E. Franconi, G. De Giacoma, R.M. MacGregor, W. Nutt, C.A. Welty (eds.), Proceedings of the 1998 International Workshop on Description Logics (DL98), ITC-irst Report 9805-03, 1998.

[54] J. Nanard, M. Nanard, Should anchors be typed too? An experiment with MacWeb, in: Hypertext '93, 1993, pp. 51–62.
[55] S. Bechhofer, R. Stevens, G. Ng, A. Jacoby, C. Goble, Guiding the user: an ontology driven interface, in: UIDIS, Workshop on User Interfaces to Data Intensive Systems, IEEE Computer Society, Edinburgh, 5–6 September 1999, pp. 158–161.

Sean Bechhofer is a Research Fellow in the Information Management Group at the University of Manchester. He received a B.Sc. degree in Mathematics from Bristol University in 1988 and has been a researcher in the University of Manchester Department of Computer Science since 1993, working initially with the Medical Informatics Group on the GALEN project. His main research interests are in the applications of description logics, particularly as a delivery mechanism for terminologies, semantic metadata and ontologies, and their use in indexing and retrieval. He has worked on the STARCH and TAMBIS projects, and is currently involved in the COHSE project, a collaboration with Southampton University's Intelligence Agents and Media group, which aims to improve the quality, consistency and breadth of linking of WWW documents at retrieval and authoring time through the use of ontological reasoning services.

Carole Goble's academic career has been spent at The University of Manchester UK, where she graduated in 1982 in Computing and Information Systems and became full professor in 2000. She co-leads the Information Management Group at Manchester. Her current interests are in metadata, knowledge representation and ontologies, and their use in hypermedia, information integration, intelligent user interfaces, and intelligent retrieval. She has worked in a variety of application areas, notably medical informatics, bioinformatics and conceptual hypermedia. Recent work includes intelligent retrieval of stock photography catalogues (STARCH), and the mediated transparent access to multiple biuoinformatics information sources through a single terminology (TAMBIS). Other projects include automating scientific database annotation process; ontology-driven intelligent interfaces for retrieval and data entry in scientific databases, and the visualisation of ontologies. Recently, she has been part of the OIL Consortium working on a language for representating knowledge on the web, co-developing an ontology-based open hypermedia system (COHSE), which serves as an example of the Semantic Web. She has published over 60 papers in these areas and has served on numerous programme committees including VLDB, ICDE, EDBT, ACM Multimedia, ACM Hypertext and the WWW Conference. She won the best paper award at ACM Hypertext 00.

ELSEVIER

Information Processing and Management 40 (2004) 957–972

INFORMATION PROCESSING & MANAGEMENT

www.elsevier.com/locate/infoproman

Word classification and hierarchy using co-occurrence word information

Kazuhiro Morita, El-Sayed Atlam *, Masao Fuketra, Kazuhiko Tsuda, Masaki Oono, Jun-ichi Aoe

Department of Information Science and Intelligent Systems, University of Tokushima, Tokushima 770-8506, Japan

Received 24 March 2003; accepted 25 August 2003
Available online 8 October 2003

Abstract

By the development of the computer in recent years, calculating a complex advanced processing at high speed has become possible. Moreover, a lot of linguistic knowledge is used in the natural language processing (NLP) system for improving the system. Therefore, the necessity of co-occurrence word information in the natural language processing system increases further and various researches using co-occurrence word information are done. Moreover, in the natural language processing, dictionary is necessary and indispensable because the ability of the entire system is controlled by the amount and the quality of the dictionary. In this paper, the importance of co-occurrence word information in the natural language processing system was described. The classification technique of the co-occurrence word (*receiving word*) and the *co-occurrence frequency* was described and the classified group was expressed hierarchically. Moreover, this paper proposes a technique for an automatic construction system and a complete thesaurus. Experimental test operation of this system and effectiveness of the proposal technique is verified.
© 2003 Elsevier Ltd. All rights reserved.

Keywords: Natural language processing; Co-occurrence word information; Co-occurrence frequency; Hierarchically

1. Introduction

How does the human begin collect accumulate the knowledge and what does he think, when people read or understood *natural language* (NL) documents. To replace these meaning of words with

* Corresponding author. Current address: Department of Information Science and Intelligent Systems, University of Tokushima 770-8506, Japan, permanent address: Department of Statistics and Computer Science, Faculty of Science, Tanta University, Tanta, Egypt.

E-mail addresses: atlam@is.tokushima-u.ac.jp, satlam@yahoo.com (E.-S. Atlam).

0306-4573/$ - see front matter © 2003 Elsevier Ltd. All rights reserved.
doi:10.1016/j.ipm.2003.08.009

the machine processing it is necessary to analyze the deep meaning of word that exists in the background. However, a huge and a variety information are needed for that, so it is difficult to replace these meaning with the machine processing easily. The thesaurus (control language dictionary) that is a typical knowledge representation solves this problem. In this thesaurus the word is classified and hierarchy by the meaning. So, thesaurus is important for a large amount document classified by the semantic content, but a hand-made thesauruses often have problems such as costs the labor and reflected the subjectivity caused by the manufacturer.

Recently various methods for automatically construction thesaurus (hierarchically clustering words) based on document data on the Web have been proposed (Hindle, 1990; Brwn et al., 1992; Petira, Tishby, & Lee, 1993; Tokunaga, Iwayama, & Tanaka, 1995). The realization of such an automatic construction method would make it possible to (a) save the cost of constructing a thesaurus by hand, (b) do away with the subjectivity inherent in a hand-made thesaurus (c) make it easier to adapt a *natural language processing* (NLP) system to a new domain. The lexical knowledge extracted by co-occurrence information (data) that used as important information to cancel the ambiguous syntax and polysemy. It is also used for clustering the related word meaning together.

Yokoyama and Shinichiro (1998) have proposed a method for classifying the meaning of verbs using co-occurrence word information and some of these verbs can not be classified correctly even an excellent result is comparatively obtained. Moreover, (Li & Abe, 1997) solved the problem of automatically clustering words by estimating a joint distribution over the Cartesian product of a partition of set of nouns (in general, any set of words) and a partition of set of verbs (e.g., noun set and verb set) and proposed an estimation algorithm based on *minimum description length* (MDL) principle. The MDL principle is a well-motivated and theoretically sound principle data compression that estimated from information theory and statistics. As a strategy of statistical estimation, MDL is guaranteed to be near optimal. Moreover, the efficiency processing of the description length is devised and alternately merging both word sets to make the thesaurus. However, the system made by Li and Abe (1997) still becomes practical use as only a sight evaluation is done and making only classification for words.

In this research, an automatic thesaurus is classified and hierarchy knowledge that human begin has constructed using the co-occurrence word (*receiving word*) and the co-occurrence frequency. When a word meaning distance is obtained from the co-occurrence relation, the word that co-occurs mutually with the same relation as a certain *receiving word* is assumed to be similar. Under this assumption, one group can bring the word group with the common feature by taking the similarity between them together. Moreover, the super-concept and sub-concept relation are given between groups by taking the similarity between each group. Therefore, when this system is characterized, the word is not only classified, but also making the hierarchy of them.

Section 2 explains *co-occurrence word information* in detail. Section 3 explains the outline of the system that becomes basic of this research. In section 4, the system is evaluated by the experimental results. Section 5 describes conclusion and possible future work.

2. Co-occurrence word information and storage technique

This section confirms the *co-occurrence word information* meanings by describing the outline about these *co-occurrence word information* and the dictionary that becomes basic of the NLP.

2.1. Co-occurrence word information

The utility of *co-occurrence word information* in the NLP system is extremely high. It is very important for canceling the ambiguous and the polysemy of words to improve the accuracy of the entire system. Various researches (Fukumoto & Tsuji, 1994; Kobayashi, Tokunaga, & Tanaka, 1996; Takahashi & Itabashi, 1998) are done for that using co-occurrence word information. Moreover, people have collected *co-occurrence word information* and the difference is seen by collection person's aspect. Our aim is to improve the accuracy of *co-occurrence information* though there is a limit amount of collection with a time restriction. This *co-occurrence word information* extracts from a large amount of corpus by the objective method and there are a lot of researches (Matsumoto, 1992; Morimoto, Iriguchi, & Aoe, 1993; Morita, Mochizuki, Yoshihiro, & Aoe, 1998; Yokoyama & Shinichiro, 1998) that use these results. Moreover, the information that defines meaning relation usually exists between two words is as follows:

Definition 2.1. When the related information α is defined between two basic words X and Y.
It is written: (X, Y, α)
The related information α has a variety of definition and retrieval demand as follows:

(a) *Relation between super-concept and sub-concept hierarchy* (*hierarchical relationship*)
The classification (concept hierarchy) represented by the thesaurus is a very simple knowledge representation and a very wide range of the application. A basic inference of this expression is because of the super-concept (high rank position) and the sub-concept becoming *co-occurrence word information* such as the concept "Clothes" and "Sports shirt". Moreover, ("Country Name" and "America"), ("Country Name" and "Canada") can also possible to define as *co-occurrence word information.*

(b) *Relation between verb and noun phrase in the case structure* (*case relation* [Hirao & Matsumoto, 1994])
In the case structure's storing meaning restriction of the noun phrase to the verb is obtained. For instance, case relation (run, dog, subject) can be obtained from (dog, animal, super-concept) and (run, animal, subject). Therefore, case relation (run, animal, subject) and (run, car, subject) are obtain.

(c) *Compound word relation*
The compound word "Canadian Nationality" is invented as ("Country Name" and "Nationality") also ("America" and "United States of America") is a compound word.

(d) *Synonym relation*
"America" and "United States of America" are synonym, also, "Cutter" and "Sports shirt" are shortening word of synonym.

Although, the surface case likes the nominative case, the objective, and the possessive, etc. can be easily decided from a large amount of corpus at the present stage. The deep case that thinks what role other words have for the verb cannot be extracted as related information. Therefore, an enough analysis cannot be done in this present structural analysis case. Moreover, Definition 2.1 is enhanced from the consideration of the frequency data of the co-occurrence (co-occurrence frequency) which define as follows:

Definition 2.2. Related information α is defined between two words X, Y, and *co-occurrence frequency* f as follows:

$$(X, Y, \alpha, f)$$

Example 2.1. The relation ("Chirp", "Bird","2") is the relation between the subject and the predicate ("Bird" and "Chirp") with *co-occurrence frequency* 2.

In this example, the word "Bird" is called a *lying word* and the word "Chirp" is called a *receiving word*. Moreover, related information (particle) is called co-occurrence relation labels.

2.2. Dictionary constructing method

The linguistic knowledge is very important in NLP system and does not limit to co-occurrence information. To make the computer analyze and generate the language, it is necessary to save the linguistic knowledge, therefore computer can be able to use it.

In NLP when analyzing the sentence and generating it, knowledge concerning the language must be brought together as a grammatical rule collection dictionary. If the word that not found in the dictionary is used in NLP system, the information in the dictionary is contradicted (i.e., an unabashedly wrong interpretation is done) and the judgment is stopped. Therefore, taking out the linguistic knowledge of the dictionary system becomes the most basic operation needed at all stages of NLP. The ability of NLP is controlled greatly by the amount of the dictionary composition and the quality.

Because the linguistic knowledge has a various classifications method, so dictionary is made at each classification. Also, the NLP system selects the plural and uses a necessary dictionary. Moreover, the collocation dictionary that brings *co-occurrence word information* together is roughly importance now. A lot of researches (Koyama & Aoe, 1995) on the dictionary constructing method and the dictionary system are done.

3. Classification and hierarchy of word

3.1. System overview

Fig. 1 shows the system overview chart in this research. This system divides roughly into three processing: (a) co-occurrence information registration (b) the word information making and (c) hierarchy.

Example 3.1. Sets of co-occurrence information are

$$\begin{aligned} C1 = \{&(\text{"}Swim\text{"}, \text{"}Young\ person\text{"}, \text{"2"}), (\text{"}Run\text{"}, \text{"}Young\ person\text{"}, \text{"7"}),\\ &(\text{"}Shout\text{"}, \text{"}Young\ person\text{"}, \text{"1"}), (\text{"}Speak\text{"}, \text{"}Young\ person\text{"}, \text{"5"}),\\ &(\text{"}Swim\text{"}, \text{"}Elderly\ person\text{"}, \text{"3"}), (\text{"}Run\text{"}, \text{"}Elderly\ person\text{"}, \text{"9"}),\\ &(\text{"}Speak\text{"}, \text{"}Elderly\ person\text{"}, \text{"2"}), (\text{"}Shout\text{"}, \text{"}Dog\text{"}, \text{"6"}),\\ &(\text{"}Speak\text{"}, \text{"}Dog\text{"}, \text{"8"}), (\text{"}Run\text{"}, \text{"}Car\text{"}, \text{"15"})\} \end{aligned}$$

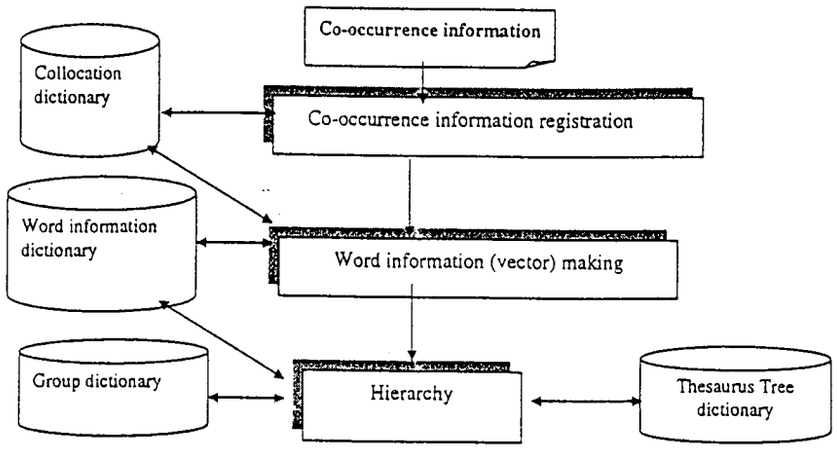


Fig. 1. System overview chart.

3.1.1. Co-occurrence information registration

In the co-occurrence information registration: *receiving word* and *lying word* are registered as co-occurrence information.

When you register co-occurrence information as in the Example 3.1 in the link trie, the distinction between the *receiving word* and *lying word* does not attach because two words are registered. Then, identifier (object-marker) * of the *lying word* is inserted in the head of the key. For instance, the co-occurrence information in Example 3.1 ("Swim", "Young person", "2") are registered by "**Young person*" and "Swim" as keys.

3.1.2. Word information making

Word information is made collectively by comparing information that shows the feature of the word mutually in some shape. In this research, to store word information the inverted index method and the vector method are expressed as a data method. As the reason, it is given to treat the vector method by the set theory, so it is easy in the programming to take AND and OR operator. Moreover, time and processing that hangs comparing will be reduced because it can have word information by the bit row. Thereafter, word information on the vector form which is composed by *receiving word* and *co-occurrence frequency* is called *Word vector*.

3.1.3. Hierarchy

It easily explains the flow of processing in *hierarchy*. First of all, the vector which is made by the *word information* is investigated to which node is similar. As a search procedure, the vector group which is a vector for the object node is made and the similarity measurement for Word vector is measured. When the tree search is finished it contains node N that have highest similarity measurement or an addition sub-concept node (called the child node) to the node N. The addition leads to the construction of the tree and the high similarity retaining becomes possible between the super-concept and the sub-concept nodes.

3.1.4. Dictionary composition

The system in this research is composed by four system dictionaries: *collocation dictionary*, *word information dictionary*, *group dictionary* and *thesaurus tree dictionary*.

(1) *Collocation dictionary* is the management of the co-occurrence word information by link tri structure,
(2) *Word information dictionary* (temporary dictionary) is the management of the Word vector. When all processing ends by the system this dictionary deleted,
(3) *Group dictionary* is the management of the Group vector,
(4) *Thesaurus tree dictionary* is the management of relationships between nodes and leaf node in the tree.

3.2. Vector information

3.2.1. Vector word information

Word information vector form is made according to the *receiving word* and the *co-occurrence frequency*. Word information vector *W_VEC* of a word with the *receiving word* of t times appearance in document can be shown as follows:

$$W_VEC = \sum_{i=1}^{t} p_i V_i \tag{1}$$

where p_i is *co-occurrence frequency*, and V_i is a vector corresponding to the *receiving word*.

All elements of the vector V_i are symbol string of 0 or 1. To add the vector of a new receiving word is just to adjust one value of symbol string that are changed from 0 to 1 newly, at the same time a linearly independent vector is added newly to the current vector.

Word information dictionary is not an important dictionary for this system because the Word vector can be made at any time based on the co-occurrence relation. However, when doing classification and clustering, this dictionary is assumed as a temporary dictionary and its information becomes important. Therefore, the *word information dictionary* is deleted with all processing terminations.

3.2.2. Group vector

The *Group* vector shows the feature of the word group and this *Group* vector is made from the *Word* vector of two or more words that belong to each group. Therefore, the word group of each group can be the group of words with same feature. Moreover, from the assumption of the Word vector, the *Group* vector is same as the Word vector when the basic group is contained only one word.

The *Group* vector information G of the *receiving word* of s times is shown as

$$G_VEC = \sum_{i=1}^{s} q_i V_i \tag{2}$$

where q_i the sum of the total *co-occurrence frequency*, and V_i is a vector corresponding to the *receiving word*.

The difference between expression (2) and expression (1) is in the point that two or more words exist in the group. It becomes the co-occurrence relation to all words with compositions *co-occurrence frequency* of them.

3.3. Similarity measurement

The word is classified by using *Word* vector *W_VEC*. The Word vector and the similarity measurement with the vector group to each node are calculated. The cosine angle of two vectors is used in this research though the similarity measurement of two vectors can be defined in various shape in vector space as

$$x \cdot y = |x||y| \cos \alpha \tag{3}$$

where $|x|$ is the length of vector and α is the angle that this vector does. When such a similarity measurement is adopted (Atlam, Fuketa, Morita, & Aoe, 2000; Atlam, Fuketa, Morita, & Aoe, in press; Fuketa, Lee, Tsuji, Okada, & Aoe, 2000; Zhang & Rasmussen, 2001). The similarity measurement of *Group* vector *G_VEC* of *Word* vector *W_VEC* becomes:

$$\mathrm{sim}(W_VEC, G_VEC) = W_VEC \cdot G_VEC = \sum_{i,j=1}^{t} p_i q_j V_i \cdot V_j \tag{4}$$

Vector V corresponding to the key word of t times, it is assumed to be orthogonal respectively:

$$V_i \cdot V_j = \begin{cases} 0 & \text{if } i \neq j \\ 1 & \text{if } i = j \end{cases}$$

Therefore, $\mathrm{sim}(W_VEC, G_VEC)$ is simplified like the following expression:

$$\mathrm{sim}(W_VEC, G_VEC) = \sum_{i=1}^{t} p_i q_i \tag{5}$$

However, because inner product is influenced by the magnitude of a vector, the similarity measurement based on this inner product normalized by using the cosine angle of these two vectors and the normalized similarity measurement $\mathrm{sim}'(W_VEC, G_VEC)$ will be

$$\mathrm{sim}'(W_VEC, G_VEC) = \frac{\mathrm{sim}(W_VEC, G_VEC)}{|W_VEC||G_VEC|} \tag{6}$$

The similarity measurement in which word information and node information were compared using vector is defined above. Moreover, similarity measurement uses the cosine that *Word* vector *W_VEC* and *Group* vector *G_VEC* of the group are similar.

Example 3.2. *Word* vector to word "Elderly person" of Example 3.1 is

$$W_VEC = (3, 9, 0, 2)$$

Group vector composed only of word "Young person" is

$$G_VEC = (2, 7, 1, 5)$$

The similarity measurement of *W_VEC* and *G_VEC* by expression (6) is

$$\mathrm{sim}'(W_VEC, G_VEC) = \frac{2*3+7*9+5*2+1*0}{\sqrt{2^2+7^2+5^2+1^2}*\sqrt{3^2+9^2+2^2}} \cong 0.92$$

4. Experiments and evaluation

In this section the method of automating construction thesaurus, the preliminary experiment for testing program and the effectiveness of this system is proposed. Moreover, in the actual experiment small-scale knowledge is initially constructed in and the system verified whether an accurate classification could be done.

4.1. Preparation

The system is constructing with addition of *co-occurrence information* on a constant amount two or more times. The number of system execution to make the classification tree correctly is important and depending on *co-occurrence information*. By some judgments it is necessary to terminate the execution automatically. To solve this problem the *stability of system* will be used. Also, it is possible that a word moves to various groups. However, if the feature of the group is specific and a certain word classified into this group then this word does not move to other groups (i.e., it is stagnated to this group). Therefore, if all words are stagnating the classification tree will be steady. So, the stagnation level that expressed the degree of the stagnation numerically is introduced as follows:

Definition 4.1

$$\text{Stagnation level} = \frac{\text{The number of word which their nodes not move}}{\text{The total number of target words classified by system}} \times 100[\%]$$

The stagnation level means the ratio in which the node is not moved even if the change in a new *receiving word* and frequency occur to the word (*lying word*). Theoretically, the stagnation level increases when the number of system execution increases.

4.2. Preliminary experiment

4.2.1. Observed data

Table 1 shows the data used by the preliminary experiment. In this experiment "feature number" is a number of *the receiving word* and the sign "<>" means the word concept.

In Table 1(a, b, c) *co-occurrence information* related to "Vehicle", "Facilities" and "Musical Instruments" is used. The co-occurrence relation of sub-concepts "Land Vehicle" and "Ship" of the super-concept "Vehicle" is prepared which is used as a *lying word*. Moreover, the words exist in these sub-concepts are used also as *co-occurrence information* for the evaluation. All these

Table 1
Observed data

Concept name	Feature number	Number of words exist in concepts	Example of word
(a) Co-occurrence information related to "Vehicle"			
<Vehicle>	217	35	Vehicle, international flight, and round trip mail
<Land Vehicle>	737	187	Curves, and private car
<Four Wheels>	22	49	Light tiger, benz, and wagon
<Two Wheels>	6	12	Motorcycle and motor-cycle under 50 cc
<Train and Bus>	71	46	Bus, streetcar, and subway
<Route Name>	4	13	Yamanote line, Chuou line, and Nanbu line
<Sky Vehicle>	735	108	UFO and helicopter
<Ship>	988	226	Aegis destroyer and Yacht
(b) Co-occurrence information related to "Facilities"			
<Facilities>	1139	615	Shrine, home, and studio
<School>	500	203	Seed School and School
<Subject>	27	70	Departments of Medicine, Engineering and English
<Faculty>	34	12	Faculties of Medicine, Engineering and Pharmacy
(c) Co-occurrence information related to "Musical instruments"			
<Musical Instruments>	86	1	Musical Instruments
<Wind Instrument>	123	37	Saxophone and flute
<Stringed Instrument>	124	37	Harp, organ, and keyboard
<Percussion Instrument>	117	35	Drum, tabor, and xylophone

Co-occurrence information prepared by human begin. The co-occurrence frequency is given at random by the range from 0 to 9.

4.2.2. Judgment method of automatically system terminating

Fig. 2 shows the transition of the stagnation level with every time executing of the system. The stagnation level increase to some degree with all co-occurrence information of the concepts and then decrease except for "Musical Instruments" because new nodes appear and the movement of the word to various nodes of sub-concepts become large, so their classifications become not clear to the tree.

From the transition in Fig. 2 the judgment method of automatically ending evaluation using the stagnation level is applied. The threshold of the stagnation level cannot uniquely decide because the stagnation level based on *co-occurrence information*. Therefore, the definition of an automatic end of evaluation is

Definition 4.2. Evaluation terminated when the stagnation level decreases than 5% or reaches 100%.

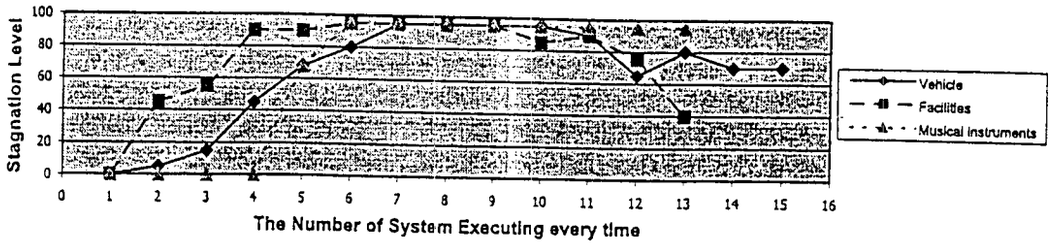


Fig. 2. Transition of stagnation level.

According to Fig. 2 and Definition 4.2 the execution is terminated after 11th times for the concept "Vehicle" and the stagnation level becomes 94.4%. However, the stagnation level is high this not mean that the classification tree is classified correctly. Therefore, it is necessary to examine the constructed system closely to confirm the effectiveness of the stagnation level.

4.3. Experiment method

4.3.1. Reserve knowledge

In the actual experiment, it used by the preliminary experiment in the foregoing paragraph. *Co-occurrence information* related to "Facilities" and "Vehicle" is given roughly the rudders as an advanced knowledge as shown in Fig. 3.

In Fig. 3 the "Vehicle" is summarized all concepts in Table 1(a) and "Facilities" is summarized all concepts in Table 1(b). Moreover, the *receiving words* "gets on", "possible to get on", and "can get on" for "Vehicle" appear 10 times, Also, *receiving words*: "build", "possible to build" and "can build" for word "Facilities" appear 10 times.

The meaning to give the advance knowledge is in the confirmation whether or not is possible to classify its accurately. In the actual experiment, when adding *co-occurrence information* the hierarchy of the sub-concept to the super-concept "Vehicle" and "Facilities" is classified with the common concepts. Moreover, the word with a different meaning is not a sub-concept of these super-concepts, and the tree is constructed. After this advance knowledge is given, the thesaurus tree is classified in the actual experiment.

4.3.2. Content of experiments

In the actual experiment, to verify the efficiency of the new method, about (one million pairs) of co-occurrence information from a data set of 25 Newsgroups from *CNN Web Site* (1995–2002)

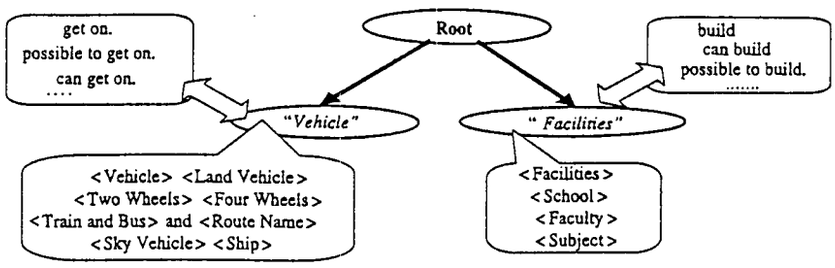


Fig. 3. Preliminary knowledge.

are used. There were various topics related to *sports*, *computers*, *politics*, *economics*, etc. As an experiment method, the above-mentioned *co-occurrence information* is added from the preliminary knowledge system of Fig. 3.

4.4. Experiment result and consideration

The changing in the system by "Vehicle" relation is shown in Figs. 4–6 as a preliminary experiment results. Fig. 4 shows the construction systems in the first time evaluation, Fig. 5 shows the second evaluation and Fig. 6 shows the construction systems after 11th times evaluation. The table that accompanies each node of the concept of the word and the frequency were shown from Figs. 4–6. Sign *A* to *M* are used as identification of nodes and assumed to be common. Moreover, in Fig. 6 each node of character string written inside is the one that classified word groups.

From Table 1(a) the relation between the concept of the word in the high rank (super-concept) and the subordinate position (sub-concept) is clear. Therefore, Fig. 6 shows the system made by

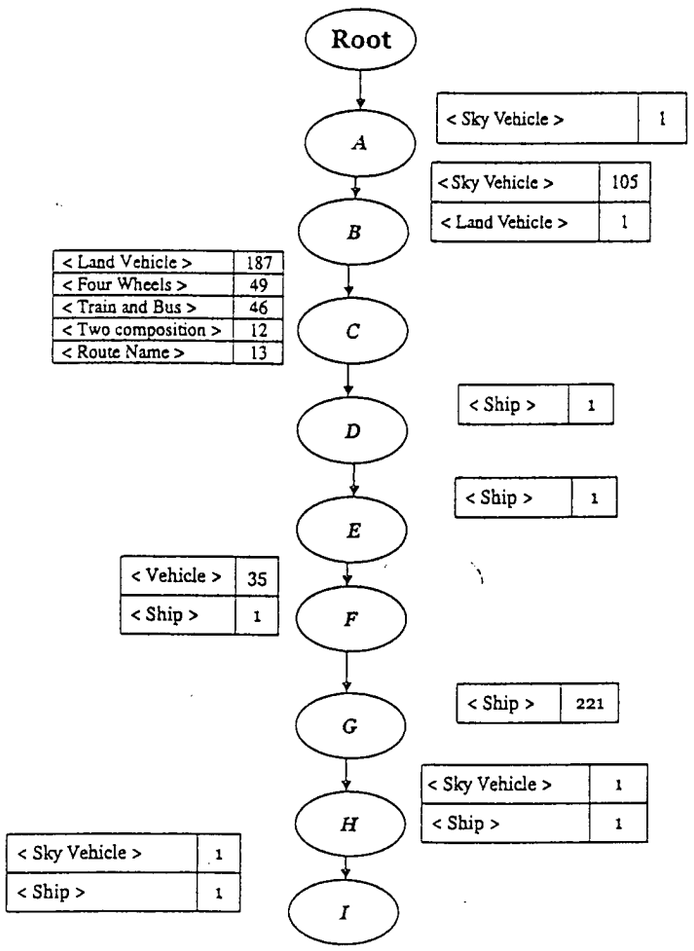


Fig. 4. Construction tree after the first time of executing.

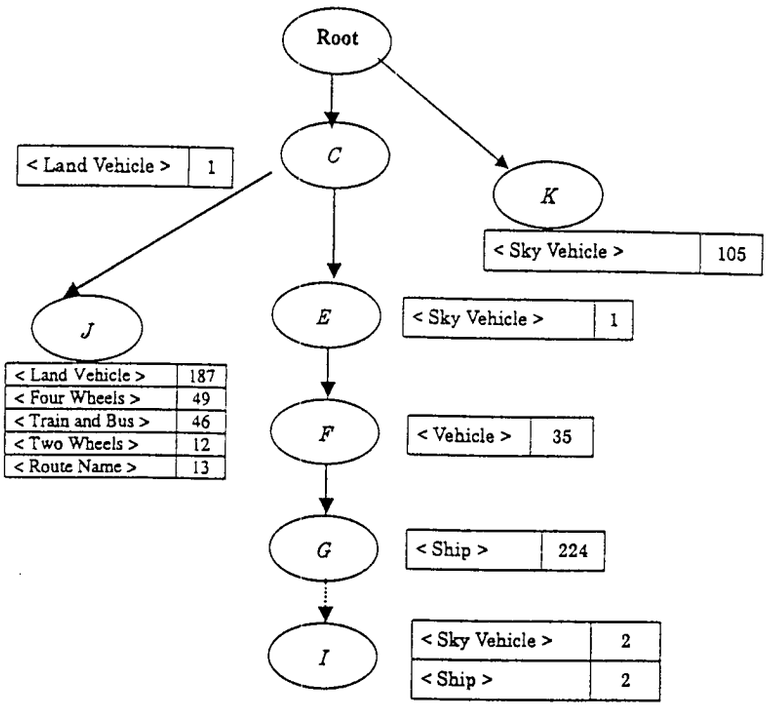


Fig. 5. Construction tree after the second times.

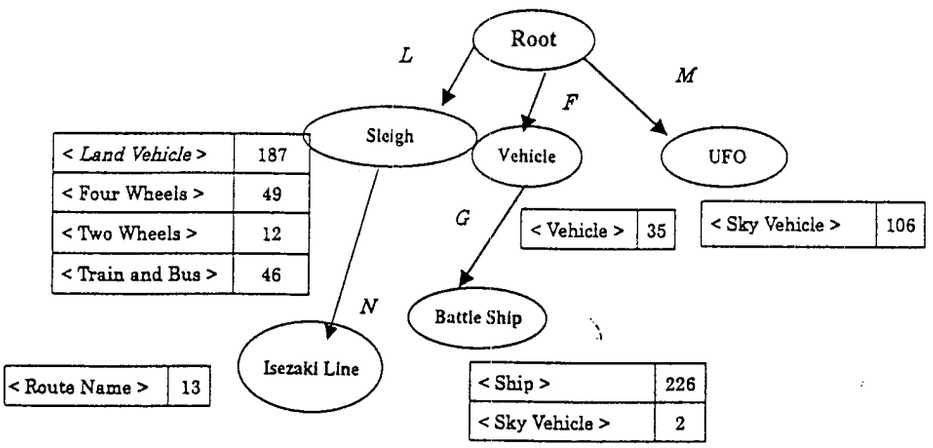


Fig. 6. Construction tree after 11th times.

hierarchically considering eight concepts related to "Vehicle". Thus, Fig. 7 is the advanced one which having the perfect classification tree. The evaluation is repeatedly by individual nodes that has distributed and then these nodes are settled as shown from Figs. 4–6. Fig. 6 can confirm the change in the system by 11 times because study is ended using Definition 4.2.

In each node of Fig. 6, node *M* represents the concept "Sky Vehicle" and node *G* represents the concept "Ship" etc. The word group with the meaning concentrates respectively is settled. It can

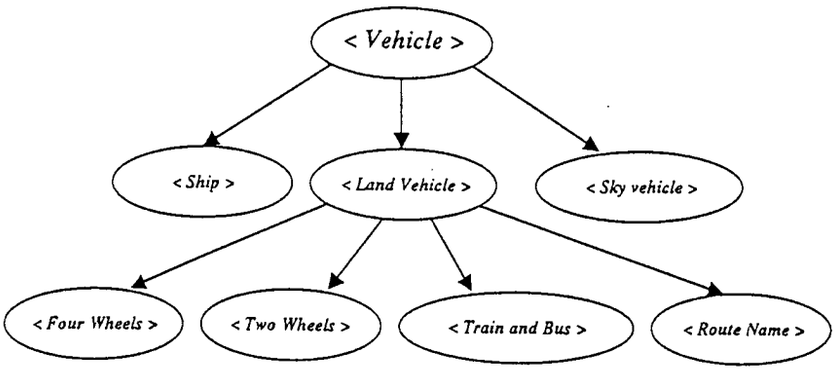


Fig. 7. Perfect classification tree.

be said that it will be classified accuracy. Two words that belong to the concept "Sky Vehicle" have been absorbed to node *G*. These words are "Hovercraft", "Spaceship" which also belongs to both concepts "Sky Vehicle" and "Ship". Therefore, it is also possible to be absorbed them to node *M* and finally they absorbed to node *G* at the 11th times of the evaluation. Moreover, node *L* is settled to one big group not being subdivided because the common characteristics of these three concepts "Four Wheels", "Two Wheels" and "Train and Bus" are settled to the concept "Land Vehicle" even the features numbers of these three concepts are less than the number of features of other concepts as shown in Table 1(a).

However, the word group that belongs to the concept "Route Name" is independently forming that group because the amount of its feature is small. This cause the features in the concept related to other "Vehicle" are gotten without the concept of "Route Name". Moreover, a top to bottom relation of concept "Land Vehicle" and "Route Name" is correct as compared with the correct answer system of Fig. 7.

Overall, the words are classified accurate and a perfect tree is constructed but with some differences in the super-concept and sub-concept relation. Moreover, the system execution is steady from the 7th times but the maintenance of that system is confirmed at the 11th times as in Fig. 6. Therefore, the stagnation level for judging an automatic end of study using Definition 4.2 is effective.

4.5. Evaluations

A part of the system constructed with the actual experiment is separately shown in Fig. 8 after nine times of evaluation. The stagnation level is 54.2%. Fig. 8 gets questions classified as accurately as sub-concept nodes as understood. For example, words related to "Baseball" have gathered in node *A* and words related to "Man" have gathered in node *D*. Moreover, the relation between node *E* and node *F* is understood easily. In a top and bottom relation of the whole, there are a few senses of compatibility group from similarity like the relation between node *B* and node *C* seems to be high and is located in a short distance.

The stagnation level studied in the actual experiment ends at 54.2%. Therefore, it is possible to understand for human even if study is automatically terminated by Definition 4.2

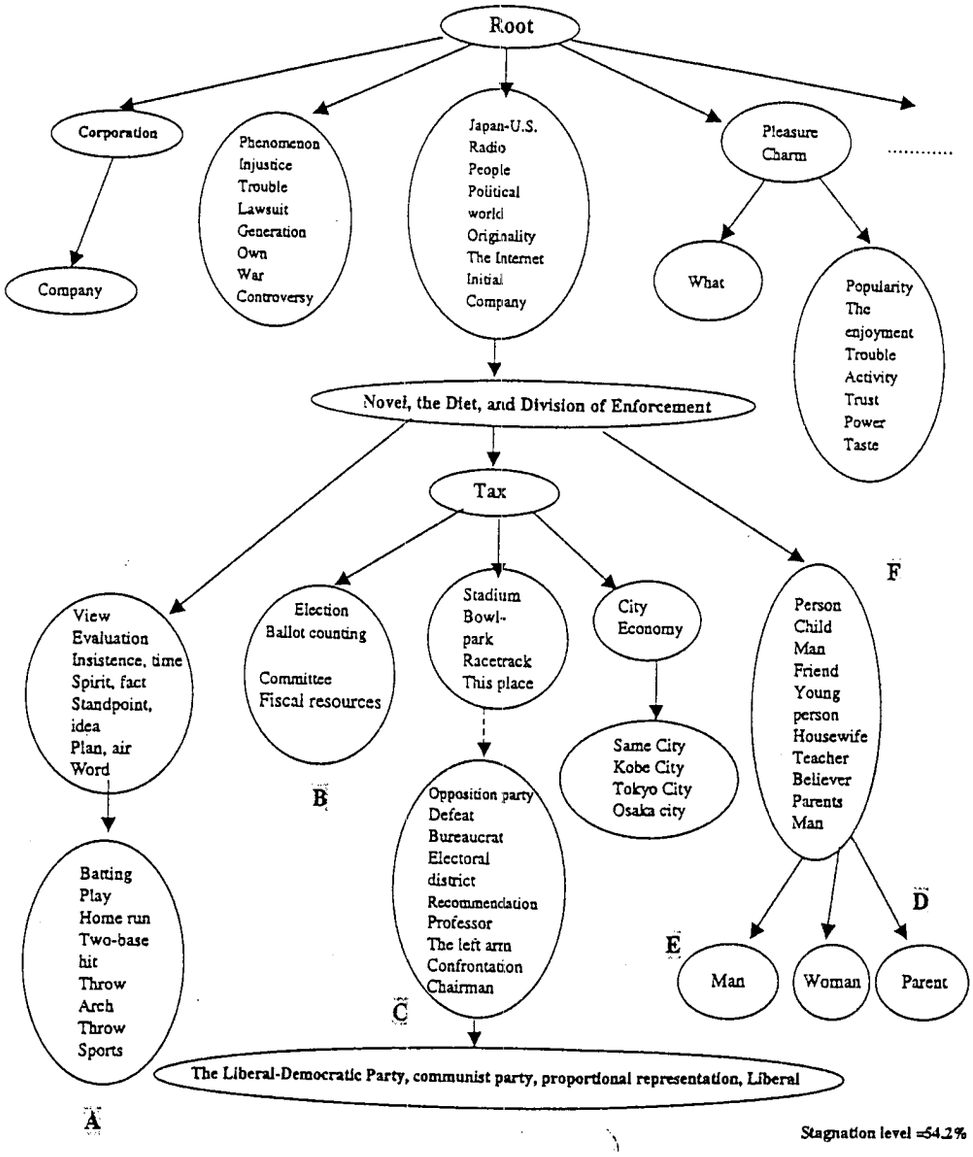


Fig. 8. Tree constructed by new method.

that the system with knowledge cannot be necessarily constructed. The factor that a system near to the correct answer system could be constructed with the preliminary experiment because the stagnation level is high. However, the result of the actual experiment described in this paper not bad because the limit of the classification according to *co-occurrence information* is used. Finally, in the actual experiment a rough limited classification can been done because the number of features of each word is limited. Therefore, it will be necessary to aim a clustering that uses larger-scale *co-occurrence information* in the future. Moreover, *co-occurrence information* is originally added by a couple, so ending the system construction

automatically based on the condition of Definition 4.2 and the complete thesaurus have been constructed.

5. Conclusion

In this paper, the importance of *co-occurrence information* in the NLP system, the classification technique of the co-occurrence word and *co-occurrence frequency* are described. The classified group expressed hierarchically and proposed the technique for constructing the system. Experimental test operation of this system and effectiveness of the proposal technique are examined.

In the future work, *co-occurrence information* should improve the system aiming by adding more each couple and use *co-occurrence information* on a large scale. Moreover, it is necessary to make automatic acquisition of the corpus by using the search engine on *Web* to prepare *co-occurrence information* on a large scale. Also, it is necessary to design the speed-up of the system construction by increasing the number of nodes.

References

Atlam, E.-S., Fuketa, M., Morita, K., & Aoe, J. (2000). Similarity measurement using term negative weight to word similarity. *Information Processing & Management Journal, 36*, 717–736.

Atlam, E.-S., Fuketa, M., Morita, K., & Aoe, J. (in press). Document similarity measurement using field association term. *Information Processing & Management Journal.*

Fuketa, M., Lee, S., Tsuji, T., Okada, M., & Aoe, J. (2000). A document classification method by using field association words. *Information Science Journal, 126*, 57–70.

Fukumoto, F., & Tsuji, J. (1994). *Cancellation of polysemy of verb based on corpus*, NLC94-24. Electronic Information Communication Society technology research report (pp. 15–22).

Hindle, D. (1990). Noun classification from predicate-argument structures. In *Proceedings of the 29th annual meeting of the Association for Computational Linguistics* (pp. 229–236).

Hirao, K., & Matsumoto, Y. (1994). *Case frame acquisition of verb from corpus and clustering of noun*, NL104-11. Information Processing Society of Japan research report (pp. 79–86).

Kobayashi, Y., Tokunaga, K., & Tanaka, H. (1996). Analysis of compound noun that uses meaning co-occurrence information between nouns. *Natural Language Processing, 3*(1), 29–43.

Koyama, M., & Aoe, J. (1995). High speed searching algorithm of hierarchy concept dictionary. In *The 51st national athletic meeting, Vol. 7 E-2* (pp. 4-235–4-236).

Li, H., & Abe, N. (1997). Clustering words with the MDL principle. *Journal of Natural Language Processing, 4*(2), 71–87.

Matsumoto, L. (1992). Rule related to the co-occurrence between words that use analysis tree data base. *Electronic Information Communication Society Thesis Magazine, J75-D-5*(3), 589–600.

Morimoto, K., Iriguchi, H., & Aoe, J. (1993). A retrieval algorithm of dictionaries by using two trie structures. *Transaaction of the IEICE, J76-D-II*(11), 2374–2383, in Japanese.

Morita, K., Mochizuki, H., Yoshihiro, Y., & Aoe, J. (1998). Efficient retrieval algorithm of co-occurrence information that uses trie structure. *Information Processing Society of Japan Thesis Magazine, 39*(9), 2563–2571.

Petira, F., Tishby, N., & Lee, L. (1993). Distributional clustering of English words. In *Proceedings of the 31st annual meeting of the Association for Computational Linguistics* (pp. 183–190).

Takahashi, N., & Itabashi, S. (1998). *Morphological analysis using word co-occurrence frequency*, 88-NL-69. Information Processing Society of Japan Natural Language Processing Society (pp. 1–8).

Tokunaga, T., Iwayama, M., & Tanaka, H. (1995). Automatic thesaurus construction based-on grammatical relations. In *Proceedings of IJCAI'95*.

Yokoyama, H., & Shinichiro, O. (1998). *Classification of meaning of Japanese verb that uses co-occurrence information from corpus*, NLC97-55. Electronic Information Communication Society (pp. 1–8).

Zhang, J., & Rasmussen, E. (2001). Developing a new similarity measure from two different perspectives. *Information Processing & Management Journal, 37*(2), 279–294.

Pergamon

Government Information Quarterly

Government Information Quarterly 18 (2001) 181–194

Informing and evaluating a metadata initiative: Usability and metadata studies in Minnesota's *Foundations* Project

Eileen Quam*

Information Architect, Minnesota Department of Natural Resources, 500 Lafayette Road, St. Paul, MN 55155, USA

Abstract

Minnesota's Foundations Project is a multiagency collaboration to improve access to environmental and natural resources information. The Project chose the Dublin Core metadata standard for web resources. Three studies were conducted: needs assessment, Bridges web site user interface, and usability of controlled vocabulary in Dublin Core metadata. Based on these findings and information architecture, the Project published best practice guidelines. Controlled vocabulary is important to facilitate access. This is relevant to the third study on Dublin Core metadata, which tested keyword searches of web pages to determine the effectiveness of controlled vocabulary in the Dublin Core subject tag. Central to the Best Practice Guidelines is the User Guide to Dublin Core, which offers an element-by-element understanding of the metadata schema. Current bibliographies and reports show further background work that informed the decision-making process for such important choices as metadata schema, thesaurus and thesaurus management software, search engine, and RDF/XML standards. © 2001 Elsevier Science Inc. All rights reserved.

1. Introduction to the *Foundations* Project

Anyone who has tried to find government information, especially via the Internet, knows that it can be difficult and often results in frustration rather than the retrieval of relevant information. Efforts to improve access do not always lead to success. Because of the multiple ways in which government resources can be described and made accessible online, discovery can be both time-consuming and tedious. Locator services developed to catalog electronic resources such as web pages, image databases, geospatial data, statistical and scientific data

* Tel.: +1-651-297-2341; fax: +1-651-297-4946.
E-mail address: eileen.quam@dnr.state.mn.us (E. Quam).

0740-624X/01/$ – see front matter © 2001 Elsevier Science Inc. All rights reserved.
PII: S0740-624X(01)00075-2

sets, and so forth generally become less and less effective as their scope of coverage expands. A more structured and descriptive approach to describing electronic resources through metadata can provide improved information search and retrieval.

In 1996, the Minnesota Government Information Access Council (GIAC) issued a report titled, *Digital Democracy: Citizens' Guide for Government Policy in the Information Age.* GIAC stated the basic principle that "public government information and services must be well indexed, easily navigable, and presented in a uniform fashion" (Minnesota Government Information Access Council, 1996, p.5).

The Legislative Commission on Minnesota Resources (LCMR) and the Minnesota State Legislature responded to this need by supporting a project proposed by a consortium of state agencies. The Minnesota Environmental Quality Board consists of 13 state agencies with responsibilities for creating and disseminating environmental information, and the Board proposed *Foundations for Integrated Access to Environmental and Natural Resource Information* to improve citizen access to environmental information.

The goal of the *Foundations* Project is to provide effective tools for citizens to discover the environmental and natural resource information they need, and to integrate access to diverse information resource types across multiple domains. *Foundations* is a collaborative project of state environmental and natural resource agencies aimed at improving public access to the data and information for which they are stewards with the following specific objectives:

- Identify citizen-oriented needs for environmental information and data maintained by or available through state agencies.
- Design an approach for integrated access to environmental information and data consistent with emerging technology standards and best practices.
- Fully evaluate the effectiveness of the design for integrated access, using external reviewers representing a diverse range of public points of view.
- Design, evaluate and implement guidelines for documentation, thesaurus and metadata implementation to promote responsive search and retrieval procedures.
- Provide skilled assistance to state agencies to catalog environmental information.
- Prepare a workable set of best practice guidelines for improving discovery and retrieval of environmental information resources based on the findings of research and testing throughout the project.
- Conduct outreach to inform key user groups and to promote implementation of the blueprint throughout state government.

The first step was to establish a project team to implement and manage the *Foundations* Project. Originally housed at Minnesota Planning, the Project moved to the Minnesota Department of Natural Resources. The project team included a project coordinator, metadata specialist, and several metadata catalogers. Early on the team recognized the need to research the current state of metadata applications. During the project, the team conducted a series of studies to inform the design and evaluate the outcomes. This paper describes the studies and their importance to the *Foundations* Project and more generally to any metadata initiative. A report, *State of Minnesota Best Practice Guidelines for Web Metadata,* contains the details of the project and the associated studies (Quam, 2000).

Table 1
Dublin Core Metadata Element Set

Content	Intellectual Property	Instantiation
COVERAGE	CONTRIBUTOR	DATE
DESCRIPTION	CREATOR	FORMAT
TYPE	PUBLISHER	IDENTIFIER
RELATION	RIGHTS	LANGUAGE

2. Architecture and components

The web presence for the *Foundations* Project is Bridges: Minnesota's Gateway to Environmental Information <http://bridges.state.mn.us>. The term "bridges" signifies the metaphoric crossing of agency boundaries and the resulting access to a wide range of environmental and natural resources information from state agencies. The gateway provides an easy-to-use search interface. After the user types in one or more search terms, the system responds with a results list of potentially relevant web resources maintained by state agencies.

2.1. Search engine

Choosing a search engine was among the first tasks of the Project. Testing commenced on three, and Ultraseek (now Inktomi) was chosen. Initially, the license covered only the Bridges website. The *Foundations* staff worked with Minnesota's Office of Technology to expand the license thus providing the State of Minnesota with a state-wide search engine that spidered all agency websites. The search engine spiders all formats of documents made available from the agency websites. It is also tuned to weight information in metadata tags for increased relevance ranking in the search results. The metadata is commonly embedded in the <head> portion of a web page, or placed in an html pointer file for pages that do not contain the <head> portion. The Bridges interface became a subset of the state search engine, North Star Search <http://search.state.mn.us>.

2.2. Metadata scheme

Dublin Core Metadata Scheme (DCMS) is the chosen scheme (Dublin Core Metadata Initiative, 1999). Bridges uses qualified DCMS that allows modifiers, schemes, and web links to be associated with a given element. Table 1 lists the DCMS 15 core elements organized according to function.

2.3. Scope of materials

The metadata can describe any document format on the websites, including HTML PDF, images, datasets, and so forth. HTML documents are easiest to describe with metadata because the metadata is embedded in the document. Bridges also has the capability of

<meta name="**DC.Title**" content="*Minnesota DNR - public recreation information map*">
<meta name="**DC.Subject**" scheme="LIV-MN" content="*Maps -- Minnesota*">
<LINK REL=SCHEMA.DC HREF="http://www.bridges.state.mn.us/servlet/lexico">
<meta name="**DC.Subject**" scheme="LIV-MN" content="*State parks -- Minnesota*">
<LINK REL=SCHEMA.DC HREF="http://www.bridges.state.mn.us/servlet/lexico">
<meta name="**DC.Subject**" scheme="LIV-MN" content="*Forests -- Minnesota*">
<LINK REL=SCHEMA.DC HREF="http://www.bridges.state.mn.us/servlet/lexico">
<meta name="**DC.Subject**" scheme="LIV-MN" content="*Wilderness areas -- Minnesota*">
<LINK REL=SCHEMA.DC HREF="http://www.bridges.state.mn.us/servlet/lexico">
<meta name="**DC.Subject**" scheme="AACR2" content="*Natural areas -- Minnesota*">
<LINK REL=SCHEMA.DC HREF="http://www.bridges.state.mn.us/servlet/lexico">
<meta name="**DC.Description**" content="*Public Recreation Information Maps (PRIM) bring together the most up-to-date information on federal, state and county lands and their recreational facilities in Minnesota. Each map displays parks, forests, scientific and natural areas, waterfowl production areas and wildlife management areas.*">
<meta name="**DC.Creator.CorporateName**" scheme="AACR2" content="*Minnesota Dept. of Natural Resources. Bureau of Engineering*">
<meta name="**DC.Publisher.CorporateName**" scheme="AACR2" content="*Minnesota Dept. of Natural Resources*">
<meta name="**DC.Contributor.PageDesigner**" scheme="AACR2" content="*Pellinen, Brent*">
<meta name="**DC.Date.Creation**" scheme="ISO 8601" content="*1998-02-03*">
<meta name="**DC.Type**" content="*Text*">
<meta name="**DC.Format**" scheme="HTML" content="*text/html*">
<meta name="**DC.Identifier**" scheme="URL" content=
"*http://www.dnr.state.mn.us/engineering/prim/index.html*">
<meta name="**DC.Language**" scheme="ISO639-1" content="*en*">
<LINK REL=SCHEMA.dc HREF="http://purl.org/metadata/dublin_core_elements">

Fig. 1. A Dublin Core Metadata Record Example

creating pointer files to non-HTML documents such as PDF files. The pointer files are small HTML documents that contain the metadata and provide a pointer to the web resource described. The advantages of making pointer files are that they will emerge higher in relevance ranking and provide a clean title and description in the results listing.

2.4. Procedures for creating the metadata

The *Minnesota Metadata Guidelines for Dublin Core Metadata: Training Manual* <http://bridges.state.mn.us/bestprac/training.pdf> details Bridges metadata creation procedures. Fig. 1 offers an example of a Dublin Core metadata record in HTML syntax. The Dublin Core elements are in bold, and the content of the element is in italics. Qualifiers, including modifiers, web links, and schemes, are also noted in the coding.

The use of standards, including international standards where possible, for data content bring consistency in applying Dublin Core elements. The example in Fig. 1 shows that the DC.Language and DC.Date elements reference the International Organization for Standardization (ISO) standards for formatting. Search engine preferences as well as library cataloging practice informed the punctuation and capitalization guidelines. When creating metadata

records, the guidelines stress the importance of consistency, ease of searching, and ease of application.

2.5. Granularity

All documents on a website do not need nor should have metadata. The goal of metadata is to get users to the most relevant documents on a specific topic. Often that topic is found on an index.html page, or one that is deeper in the website, with links to allow further exploration of that topic. Metadata typically belongs on the index page, and the user can navigate deeper into the site if more specificity is desired. Web-based periodicals offer another example for metadata and granularity. In this case, the goal is to get the user to the main page of the periodical, which may be a template containing the latest issue or an opening page that links to the archives. As each new issue is published, the metadata remains constant on the opening page or the template page with no need to add metadata for each subsequent issue.

Because of these rules of thumb regarding granularity, Bridges has added metadata to approximately the top 50% of the Web pages spidered on the Bridges search site. This figure was calculated by tracking the number of web pages to which metadata was added compared with total number of pages on the websites.

2.6. Metatagging software

TagGen, from Hiawatha Island Software, was chosen as the Dublin Core metadata generator for the Foundations Project. Using TagGen's Dublin Core Editor, metadata can be quickly and efficiently inserted into the HTML headers of existing and newly created web pages of Minnesota state agencies. The Foundations Project worked with Hiawatha Island engineers to create a version for qualified Dublin Core that contains an editing screen with drop-down menus. The menus can be customized with qualifier values to suit the metadata applier. Due to these menus, no previous experience in adding metadata is required nor is HTML coding skill required. The application includes standard metatags plus the 15 Dublin Core elements. Screen shots of TagGen showing typical application of qualifiers for the 15 elements are included in the *User Guide for Dublin Core Metadata* that is contained in the *Best Practice Guidelines* (Quam, 2000).

Another feature of TagGen is its lack of interference or conflict with previously placed metatags, page design, or setup. TagGen can be mounted on Windows 98, 2000, or NT platforms, offers multiple website support, and gives the user the ability to do rapid updates of many pages simultaneously. Other metatagging software considered is listed in the Metadata Software Report in the *Best Practice Guidelines* (Quam, 2000).

2.7. Vocabulary control/thesaurus

To bring consistency in the words used to describe the resources, the *Foundations* Project implemented two types of controlled vocabularies: a thesaurus, and the Content Classification Engine. The purpose of a controlled vocabulary is to bring together like concepts,

```
<meta name="DC.Subject" scheme="LIV-MN" content="State parks -- Minnesota">
<LINK REL=SCHEMA.DC HREF="http://www.bridges.state.mn.us/servlet/lexico">
```

Fig. 2. Dublin Core Scheme Qualifier Example

synonyms, and related terminology. The controlled vocabulary in the form of a thesaurus can assist the user in selecting appropriate search terms the searcher may not have envisioned at the onset. Such assistance takes the form of two types of cross-references:

- *See references* lead users from a synonym to the preferred terms or phrase for a concept, or from variant spellings of a name or title;
- *See also references* suggest related concepts that are not precisely synonyms yet bear a close connection to the chosen term.

A true thesaurus also provides hierarchical views among terms, allowing for broader and narrower ordering of concepts. A thesaurus provides a worldview on a particular domain or discipline by organizing concepts and subject matter, and creating a comprehensive, logical understanding of the entire knowledge area. A further advantage of a thesaurus is the ability to share the vocabulary. If other similar projects use the same vocabulary, cross-domain searching is facilitated. The *Foundations* Project started with a specific thesaurus, which is now used by other agencies within the state, and by some other state metadata efforts. Using the same set of terms for similar concepts brings coherence and consistency to searching.

Related to the notion of controlled vocabulary are naming conventions that bring about consistency in recording names of people and organizations. Name authority files and other mechanisms can prescribe a preferred form of entry. Examples include a preferred listing or code to record county names; preferred format for geographic names (Lake Superior or Superior, Lake), and preferred entries for governmental agencies and subunits. When a specific scheme of authority control is used, it is possible to reference it in the metadata, or at least in the metadata application user guide. Qualified Dublin Core metadata uses a scheme qualifier that provides a scheme name and a link to the referenced scheme. This feature offers the sophisticated searcher added information about the metadata element; metadata about the metadata as it were. Fig. 2 shows an example of the use of this feature of Dublin Core. In the example, the **scheme** for DC.subject is LIV-MN (an abbreviation for the Legislative Indexing Vocabulary, MN edition) and the web link is cited in the **link rel** location (both noted in bold).

The *Foundations* Project chose the Legislative Indexing Vocabulary (LIV), maintained by the Library of Congress and Congressional Research Service, as the primary thesaurus. The DC.Subject element contains terms from this thesaurus. Primary reasons for selecting LIV include:

- Appropriate vocabulary, including sufficient depth of terminology.
- Speed and accuracy of locating appropriate terms in the software.
- Flexibility of both the vocabulary and search capabilities.
- Editing capabilities of Lexico software.

Further information can be found in the Thesaurus Report part of the *Best Practice Guidelines for Web Metadata.*

After evaluating and choosing this thesaurus, the Project purchased it along with Lexico thesaurus management software. The Bridges version of the thesaurus is LIV-MN, indicating that the thesaurus is now Minnesota-specific. This application is an important part of vocabulary and metadata maintenance, and was an aspect in the design of the final usability study discussed below. LIV-MN is located at <http://bridges.state.mn.us/servlet/lexico>.

In mid-1998, the LIV appeared as the best choice for Bridges. The common state subject tree Dr. Jessica Milstead created for the State of Illinois did not yet exist. The "Jessica tree" <http://states.gils.net/hierarchical.pdf> is available for adoption by all state metadata efforts. See accompanying article in this issue about the creation of this subject tree and the Illinois experience.

2.8. Browseable topic categories/CCE

The other use of controlled vocabulary in the Project involves the Content Classification Engine (CCE), a Yahoo-style topic directory that is an add-on to the Ultraseek/Inktomi search engine. The full CCE hierarchy can be accessed from the State of Minnesota's North Star Search <http://search.state.mn.us>. The Bridges interface links to the Environmental and Ecology topic of the CCE.

CCE can be used in a point-and-click fashion, and the results produced rely upon rules created by the administrator. The rules tell the search engine where to gather web pages for the specific topic. Typical rules are folder-level URLs, but can also include rules that incorporate the controlled vocabulary contained in the Dublin Core subject element. As an example, consider the topic Hazardous Wastes. A great deal of information on this topic exists in a particular folder on the Minnesota Department of Agriculture website and in another folder on the Minnesota Pollution Control Agency website. These became the first two rules. However, scattered pages on the topic exist in various other agency websites including the Metropolitan Council and Minnesota Department of Natural Resources websites. These websites have Dublin Core metadata attached to the pages and utilize the LIV-MN controlled vocabulary. Therefore, the rules that direct the search engine where to pull information for the results on this topic include:

Rule 1: url:www.mda.state.mn.us/DOCS/
Rule 2: url:www.pca.state.mn.us/waste/
Rule 3: dc.subject:"Hazardous wastes"
Rule 4: dc.subject:"Hazardous waste disposal"
Rule 5: dc.subject:"Medical wastes"

This is a powerful leveraging of the Dublin Core subject element. The search engine will find web pages on the chosen topic no matter where they reside without the searcher needing to know the exact terminology.

Beyond this, the searcher can enter terms to search within those topics. Finally, CCE suggests Related Topics when a searcher enters keywords. The user interface study, however,

discovered that this was potential source of confusion for users. Changes in the placement of Related Topics have largely addressed this confusion.

The development and refinement of the Bridges components discussed above resulted from a series of studies including needs assessment, interface usability, and an assessment of adding controlled vocabulary in the subject metadata element to web resources to improve information retrieval. The following sections discuss these studies.

3. Informing the design of bridges

An important contribution of the *Foundations* Project to other metadata initiatives is the use of empirical research to inform the design of the system and assessment of the resulting product to identify the utility of a metadata approach to improve discovery of and access to government information. The project team conducted a needs assessment study, a user interface usability study, and an assessment of the use of metadata and controlled vocabulary.

3.1. Needs Assessment Study

The *Information Needs Assessment: Summary of Survey Results* summarizes information needs identified by Minnesota residents in three surveys conducted in 1997 (Foundations Project, 1998). It describes the methodology of the assessment to develop a description of "what people want to know" about Minnesota's environment and natural resources. One aspect of the study was to prioritize these information needs. The scope of information deemed relevant included all data and information collected and managed by all of the Minnesota Environmental Quality Board agencies participating in the *Foundations* Project.

The focus of the study was on citizen needs, not the information needs of experts (e.g., the data and information needs of local governments and businesses). Although state and local experts provide a critical link in meeting public demand for information about Minnesota's environment and natural resources access, the goal of Bridges is to provide citizen access to environmental data. The complexity of many environmental questions requires a level of time and expertise beyond that of most citizens. With a system such as Bridges and the access to data and information about Minnesota's environment and natural resources it provides, state and local government will be better positioned to serve their constituents.

Information needs cover a broad spectrum of data and information about Minnesota's environment and natural resources. The results of the study indicated clearly that some topics and types of information are requested more frequently. Yet it was the diversity of information that best characterizes "what people want to know." From a design perspective, the diversity and scope of information needed identified by survey respondents indicated that broad coverage of the metadata application would be necessary.

The researchers used content analysis software to tabulate the frequencies of similar terms to arrive at a preliminary list of high priority information needs. These served as a starting point for creating an access framework. The researchers then inspected the terms for common themes and

Develop (-ment, -ments)
Environment (al)
Lake (s)
Location (s)
Maps
Minnesota (-'s, MN, MN's)
Park (s)
Pesticide (-s, /herbicide)
Pollute (-s,-tion, -ants)
Quality (air, water, land, etc.)
Recycle (-ables, -ed, -ing)
State (county/local)
Sustainable (-ability)
Water
Wild (-erness, -life)

Fig. 3. Frequently Occurring Terms For User Information Needs

context, resulting in a list of common needs or themes. Several terms appeared frequently in the survey responses. Fig. 3 lists the most frequently used terms identified in the study.

An examination of terms such as "water" or "pollution" found substantial variety in the meanings of those terms reflecting specific information needs. For example:

- People want information about the interaction between environmental conditions and human health, particularly as it pertains to their community. Water quality, air quality, pollution, toxics in products and pesticide use are examples of this information need.
- People want to know the location and variety of recreational opportunities in Minnesota. Survey respondents identified information on hunting, fishing, boating and camping in each of the surveys.
- People want to information about the condition, location and trends in Minnesota's natural systems such as lakes, rivers, land, habitats, plant and animal populations.
- People want information about the costs and results of public initiatives on behalf of Minnesota's environment and natural resources. This includes such things as public spending, regulatory activity, new legislation, volunteer opportunities, and the results of public action.
- Interactions between the environment and human communities, especially in their community. Many survey respondents sought information about development and the environment, businesses and environment, agriculture and the environment, with particular concern for their local communities.
- People identified information to help them learn more about environmental issues and recent events. These information requests also included a desire for information about new laws, public initiatives, state and local policies, and background information to better understand issues in the news.

The needs assessment results informed the *Foundations* Project on two important aspects. First, it provided a basis for determining the types of information resources that citizens would like to access. Second, it provided terminology that could be used as a basis for describing the resources so users could discover them. From a metadata perspective, the crucial offering of this study is that web pages need sufficient description for retrieval to allow easy access to the important topics indicated.

3.2. User Interface Study

Foundations Project team conducted a usability study for the design of a search interface in spring 2000. The complete results of the study can be found in *Foundations Project Usability Testing: Bridges Search Interface Study* (Foundations Project, 2000a). The study's goal was to gather information on the ease and comfort level of searchers using the Bridges web interface, and to collect information on successful and unsuccessful strategies used for searching governmental and environmental information on the site. An additional goal was to determine how metadata affects search results.

Usability testing of Bridges search interface involved developing the plan, designing the necessary materials, finding participants, and choosing the testing locations. Upon completion of the testing, the researchers conducted statistical analysis on the data to find what search strategies the public at large uses. The researchers were also interested in assessing the effect of metadata in efficient retrieval. The Project used the definition of usability in the ISO 9001 standard (International Organization for Standardization, 1994): "The usability of an interface is a measure of the effectiveness, efficiency and satisfaction with which specified users can achieve specific goals in a particular environment with that interface."

The researchers first created a set of test questions that could be answered by information contained in resources on several Minnesota state agency websites. The researchers searched the resources to determine the location of answers and the level of difficulty in finding them. They also delineated criteria to evaluate successful completion of the task and identified the data to be collected, for example:

1. Did the user find the specific resource(s) that contained the correct information to answer the question?
2. What search strategy or strategies the searcher used, for example
   - Keywords
   - Boolean
   - Natural language
   - Quotation marks, truncation, and so forth
3. How many resources were retrieved and of those, how many were relevant to the question?
4. What strategies did users most often use to find answers to a question?
5. How much time did it take users to answer each question? What was the average or maximum time taken to perform the task?

Researchers used an information form to record data about the searchers, and used a stopwatch to monitor time to complete the tasks. As participants located the specific page where the answer to each question could be found, researchers transcribed the search terms and strategies used. Participants included a test searcher, two searchers from the Minnesota Department of Natural Resources familiar with web searching, one person with some experience in this type of discovery from the University of Minnesota, and two participants from the College of St. Catherine who had very little searching experience. Participants completed a brief user profile

questionnaire to assess the skill levels of participants. The following provides some general results from the study; the complete report provides specific details and results.

The key finding from the study was that all searchers but one had success in searching the Bridges site. Search times decreased as users became more familiar with the Bridges site. Each participant's average time per question was well under the five-minute limit, with minimum and maximum average times of 0.56 s to 4.16 min respectively. Most participants found the correct site promptly and easily. Only one user opted to make multiple searches for each question (this person also had the lowest success rate indicating that the problem was more of skill level and search techniques than the Bridges site itself).

When a search resulted in multiple hits (a common occurrence commented on by at least one user), most participants found the correct page near the top of the results list or by a reasonable amount of scrolling. Total number of hits per participant varied greatly, with an average low per question of 461.4 and a high of 15,607. This indicates that the addition of terms in the metadata weighted these sites toward the top of long lists of hits.

Participants used varying search strategies with keyword being the most common. (52% of searches used keywords.) They also used, but to a more limited extent, Boolean (42%), related topics suggested via the Content Classification Engine (CCE, discussed above) (4%), natural language, such as full question phrases (1%), and quoted phrases (1%).

The researchers discovered useful information by observing the searchers and from answers on the post-test questionnaires. Most searchers were comfortable with keyword searching and the more advanced Boolean features. Although Boolean searching was popular, one user had trouble with the Ultraseek search operators. None of the searchers went to the Help screen to find this information. The least experienced searcher indicated frustration with the hits received using Boolean and commented negatively on their high number. Those searchers who chiefly used keyword searching had good success rates. This was an important finding since inexperienced users are more likely to use keywords.

Opinions on using CCE were mixed: two participants ranked CCE easy to use, two were less enthusiastic and one abstained from answering. A searcher who had trouble with the wording of questions expressed interest in CCE. Use of CCE varied among the searchers. The more experienced searchers had some success using this search strategy (knowing, for instance, which Minnesota state agency would be likely to offer the necessary information), but those with less experience seemed confused by it.

Based on the results of the study, the researchers concluded that for this set of users, the Bridges search site performed very well. Most users were able to complete tasks successfully and without assistance. As a result of the study, Project staff made minor changes to the interface, such as changing the location of CCE (linked related topics) to lessen confusion, and improving access to the topic hierarchies.

3.3. Usability study on metadata and controlled vocabulary

With any metadata application, there is the question of usefulness. Is it worth the resources to add metadata to resources? Does structured metadata provide a basis for increased effectiveness of search and retrieval? The *Foundations* Project focused attention on these and other questions through its final and most important study, *Foundations Project Usability*

Testing: Dublin Core Metadata and Controlled Vocabulary (Foundations Project, 2000b). This study began before the project staff completed its final decisions on standards and content of the metadata and was ongoing throughout the implementation of the project. The goal of the study was to test keyword searches of web pages to determine the effect of controlled vocabulary use in the Dublin Core subject element (DC.Subject). Researchers identified a set of search terms for the study before the addition of metadata and used those terms consistently throughout the phases of the study. This set consisted of terms from the controlled vocabulary as well as uncontrolled keywords and synonyms. An independent analyst examined the change in rates of dead-end searches, excessive results, and irrelevant results before and after the addition of controlled vocabulary terms in the Dublin core subject metadata element.

The study comprised three separate passes

- Pass 1 occurred before any metadata had been added to the pages; all search terms were considered keywords, or uncontrolled terminology.
- Pass 2 used the same search terms (by now a mix of uncontrolled terminology and thesaurus terms) after metadata had been applied to about half of the appropriate documents.
- Pass 3 used the same search terms when the Project was very nearly complete, and metadata had been applied to nearly all appropriate documents.

A summary of spreadsheet results for the three passes is available <http://bridges.state.mn.us/user2spreadsheet.pdf>.

The researchers designed a testing methodology that included a data collection instrument to record the searches and results. The results of the study assisted project staff in assessing the success and utility of the *Foundations* metadata project. For each of the passes, the analyst recorded the following information:

- Searches that had no results ("dead-end searches").
- Searches with over fifty results.
- Searches that resulted in an insufficient number of results.
- Searches that found inappropriate sites.
- Search results using designated keywords.
- Search results using designated terms from the Legislative Indexing Vocabulary.

Several factors affected the increased search effectiveness. First, the Ultraseek search engine, spiders and robots were subject to improvements, including a capability for the use of truncated search terms, more consistency in total returned hits, and better help screens. Over time, response time improved as well, with shorter waits for downloads. Second, the agencies populated more agency websites and pages with *Foundations*-specific metadata, including the use of controlled subject terms from the LIV thesaurus.

By the third pass in the study, data recorded by the study's analyst provided evidence that a major goal of the *Foundations* Project had been achieved: describing Minnesota state agency web sites with Dublin Core metadata resulted in improved search results. Further, with the higher numbers of websites that contained metadata and included in the spidering, development of the LIV-MN thesaurus, and Ultraseek search engine's ability to use quotes

and Boolean symbols, search results improved in both quantity and accuracy. The analyst's searches were done very simply, using the prescribed list of terms that were either from LIV-MN terms or keywords. With good results from such basic and replicable searching, it is likely that both experienced and nonexpert, nontrained users would have little problem achieving abundant and accurate search results.

It is one thing to show improved search effectiveness but another to assess whether the benefits are worth the cost, specifically, the cost of adding Dublin Core metadata to web pages. The study provided evidence that there are clear benefits of adding a subject metadata element that contained controlled vocabulary terms.

The *Foundations* Project staff studied the time and cost of adding metadata. The staff determined that it takes an average 5–7 minutes per web page to add metadata. The lower time is often due to familiarity with the subject matter; for example, it takes less time to create the metadata if the creator has worked in the relevant subject area. The greater time reflects pages that are more difficult to analyze. The metadata creation time of 5–7 minutes assumes the person creating the metadata is familiar with TagGen, the metatag software, and LIV-MN, the controlled vocabulary. The average pay rate of the eight *Foundations* staff metadata catalogers was $17/hr. Assuming a 75% production rate, the cost comes out to between $1.12 and $1.50 per web page. Granularity certainly has an impact on cost as well. If metadata creators work with approximately the top 50% of pages in web folders, the addition of metadata is much more cost-effective than if each page contained metadata.

4. Best practices for web metadata

The *Foundations* Project developed an information architecture that has improved public access to environmental information within Minnesota State agencies, and greatly enhanced Internet search and retrieval accuracy. A key contribution from the *Foundations* Project is the *State of Minnesota Best Practice Guidelines for Web Metadata* (Quam, 2000). As the capstone for the project, the *Best Practice Guidelines* includes practical instructions for metadata creation as well as background documentation and data that support the recommendations and standards that have made the project so successful.

The *Best Practice Guidelines* contains the *User Guide for Dublin Core Metadata,* based on the Dublin Core Metadata Initiative's (DCMI) user guide but customized for the State of Minnesota's use of qualified Dublin Core. While this user guide contains some Minnesota-specific practices, it has been adapted by many other metadata state, city, and international efforts. In addition to giving information on application of specific Dublin Core elements and their qualifiers, the *User Guide* also offers visual cheat sheets, a glossary of relevant web terminology, information on TagGen metatag generator, and subheadings to be used with LIV thesaurus terms. The *User Guide* will be maintained for currency on an on-going basis.

Appendixes in the *Best Practices Guidelines* include full reports of the studies discussed above and a series of relevant Project background reports. It also contains comprehensive bibliographies on relevant metadata topics including, for example: searching and search Engines; GILS Metadata; and Dublin Core Metadata

The *Foundations* Project hopes the information contained in the *Best Practice Guidelines*

will assist others as they design and implement other metadata initiatives in Minnesota and elsewhere.

5. Summary

The success of the *Foundations* Project is largely due to careful research, decision-making, and application of good metadata practices. Usability testing of both the interface and the use of controlled vocabulary in the Dublin Core metadata subject element shows that these practices are worth the effort.

To be effective, metadata needs to be maintained, as does the vocabulary and its offshoots, such as CCE. A good example of this is when web page creators use an old page as a template to produce a new page on another topic. If the old metadata is not attended to, the search engine may well consider the page a duplicate of the old page, and refuse to include it. Maintaining metadata has not been as easy to promote as the initial creation of metadata. As with any system or service, ongoing training and education is necessary to maintain the gains achieved so far in the *Foundations* Project.

The State of Minnesota's Office of Technology is now recommending the *Best Practice Guidelines* and its metadata application for web pages on all Minnesota State agency sites. As the collection of electronic documents on Minnesota state agencies' Internets and Intranets grows, metadata will be integral to assist citizens to find useful information. Agencies are now investigating the use of metadata for records management and archiving. Metadata descriptions of web pages allow information to be found more easily and accurately. Metadata helps search accuracy on commercial search engines. Finally, consistent best practices allow a standardized approach to electronic document description, retrieval, or archiving.

References

Dublin Core Metadata Initiative. (1999). Dublin core metadata element set, version 1.1: Reference description. Available URL: http://dublincore.org/documents/dces/

Foundations Project. (1998, April). Foundations Project information needs assessment: summary of survey results. St. Paul, Minnesota. Minnesota Planning. Available URL: http://bridges.state.mn.us/INA.pdf.

Foundations Project. (2000a, March). Foundations Project usability testing: Bridges search interface study. St. Paul, Minnesota. Minnesota Department of Natural Resources. Available URL: http://bridges.state.mn.us/userstudy.pdf.

Foundations Project. (2000b, July). Foundations Project usability testing: Dublin Core metadata and controlled vocabulary study. St. Paul, Minnesota. Minnesota Department of Natural Resources. Available URL: http://bridges.state.mn.us/user2study.pdf.

International Organization for Standardization. (1994). ISO 9001:1994. Quality systems – Model for quality assurance in design, development, production, installation and servicing. Geneva: International Organization for Standardization.

Minnesota Government Information Access Council. (1996, December). Digital democracy: Minnesota citizens' guide for government policy. St. Paul, Minnesota. Minnesota Department of Administration.

Quam, E. (2000, August). State of Minnesota best practice guidelines for web metadata. St. Paul, Minnesota. Minnesota Department of Natural Resources. Available URL: http://bridges.state.mn.us/bestprac/index.html.

Problems of Thesaurus Construction
Amitabha Ghose; Anand S Dhawle
Journal of the American Society for Information Science (pre-1986); Jul 1977; 28, 4; ABI/INFORM Global
pg. 211

Problems of Thesaurus Construction*

The basic assumptions of the different methods of thesaurus construction are analyzed both from the epistemological point of view as well as from the information retrieval needs of the users. The construction of thesauri for social sciences becomes particularly difficult, because of the inherent ambiguity of the terms used in social sciences. Since the idea of a "growing thesaurus" is impracticable, it is suggested that an empirical study of the semantic differential of different sociological groups would suffice for the purpose of thesaurus construction.

Amitabha Ghose
Professor, Production & Quantitative Methods Area
Indian Institute of Management
Ahmedabad 380015, INDIA

Anand S. Dhawle
Librarian & Member of the Faculty
Vikram Sarabhai Library
Indian Institute of Management
Ahmedabad 380015, INDIA

• Introduction

Thesauri for different subjects like science and engineering, psychology, sociology, and political science have already been constructed by different institutions for documentation (*1-5*). The purpose of this paper is to discuss some of the problems of thesaurus construction in the methods currently discussed in the literature (*6-9*). Thesaurus construction has basically two aspects: (i) selection of keywords/phrases of the particular subject for which the thesaurus is constructed, and (ii) establishing an interrelationship among those terms which deal with the problem of synonymity and hierarchical relations among terms. The importance of both these aspects for information retrieval purpose is evident.

While reviewing the state of the art of thesaurus construction, we discuss some of the major difficulties in constructing thesauri according to two different methods of thesaurus construction: (i) *a priori* classification or faceted category method, and (ii) *a posteriori* statistical method based purely on frequency of terms. We shall discuss how the problem of interrelationship of terms in the natural sciences is quite a distinct one from that of the social sciences, primarily becuase even today the concepts used in the social sciences are not as exactly defined as they are in physics, chemistry, medicine, or engineering. We shall outline a method for tackling the problem of ambiguity in the social sciences for the construction of thesauri.

*Revised version of the paper presented at the Seminar on Thesaurus in Information Systems held at Bangalore, India, 1-5 December 1975.

© 1977 by John Wiley & Sons, Inc.

Reproduced with permission of the copyright owner. Further reproduction prohibited without permission.

• A Review of the Methodology

In traditional classification theory certain terms characterize important branches of a subject, some terms specify the important topics within the different branches, and there can be still some other terms for the different subtopics and so on. The selection of these terms and the hierarchical relations among them could be considered as a first step towards thesaurus construction. If such a thesaurus is to be of use for information seekers, then it has to be assumed that the important topics and subtopics of the subject are uniquely determined by terms which are well known to the seekers of information.

Due to the modern interdisciplinary development of knowledge it has almost become impossible to characterize an important topic just by one term (*10*). The topic may have many "facets." It may be important from different points of view. For instance, a text on quantum physics may be important not only for nuclear physics, but also for mathematics, logic, and philosophy. Similarly a text may be important for the idea of "democracy," "ethics," and "economic growth" (*11, 12*). Ranganathan has tried to solve this problem of "many facets" of a topic by his five categories of PMEST (*13, 14*). Vickery has extended the notion of faceted classification by a mere comprehensive list of fundamental categories (*15-17*). A recent version of the "category approach" for the universe of knowledge in terms of multidimensional vector spaces is given by Meincke and Atherton (*18*). However, in all these "category approaches" to thesaurus construction it is assumed that the categories are unambiguous and mutually exclusive. But such an assumption is difficult to justify semantically. In fact, one of the main problems of thesaurus construction for social sciences is the difficulty of finding key words which are unambiguous. To determine *a priori* certain fundamental categories and to base the construction of thesaurus on these may simplify the procedure, but would fall far short of satisfying the information retrieval needs of the users.

An empirical approach to the problem of selection of key words and their inter-relationship would consist in an analysis of the terms and "term clusters" used in a large number of texts on a given subject (*19*). The selection of key words or phrases for the subject is done by scanning through documents in a collection and making one list of words/phrases significant for information retrieval purposes and another list of common words (negative dictionary) which are common to most subjects and hence have no significance for information retrieval purposes of the subject under consideration. In preparing the list of significant words, it is sufficient to consider only the stems of words so that any other word dervied by the addition of a suffix to the stem word can be identified with its stem word. This process would significantly reduce the length of the list of words.

Fully automatic methods have been successfully used for such a selection procedure (*20-29*). This is based on making the computer find out the frequency of terms found in a given collection of documents, and then choosing as significant those terms for which the frequencies are neither too high nor too low (*30, 31*). One studies the relative frequencies of words in a collection of texts on a given topic. These words can now be arranged in a sequence according to the magnitude of the averages of the relative frequencies. At this point, it is *expected* that the terms important for information retrieval would cluster around the middle. Those too much to the left or the right can be considered to be insignificant for information retrieval. The decision where to fix the limits, however, remains arbitrary. The terms of very high frequency are considered as common words and those of very low frequency are considered to have little relevance for the subject. Such terms are considered to be insignificant from the point of view of information retrieval.

Statistical methods are never meant to give exact results. One should not expect that by eliminating the words with very high and low frequencies one would get just the key words which are the only significant ones. For information retrieval it is necessary to make sure that the method guarantees a high degree of probability of getting a set of key words approximating the desired one. The inference is only about a certain frequency pattern of terms for a given subject. However, it is still not possible *a priori* to make the logical inference that in this way one would get just the words which are significant for information retrieval purposes. One could at the most say that it is not a bad hypothesis. For a text on physics the frequency of terms like "love," and "beauty," will be "too low." Of course, even some important terms in physics may have very low frequencies in a sample of texts and yet be very significant for the topic. Hence the only way to test the hypothesis is to put it to use and get the feedback of the users. Thus if *most of the time* the vocabulary of the user approximately coincides with the "treasure" of key words determined by eliminating words "too low" and "too high" in frequency, then the method of eliminating high and low frequency words can be accepted for thesaurus construction.

Let us now consider how inter-relationships of terms is established by means of a term document matrix based on the frequency theory (*8, 32, 33*). The frequency of terms in each document belonging to a collection of documents on the subject for which a

Reproduced with permission of the copyright owner. Further reproduction prohibited without permission.

Let us consider a term property matrix for computer science like the one considered by Salton (*22*, p. 51):

	Abstraction	Physical Object	Hardware	Software
Computer				
System				
Program				
Machine				
Equation				
Logic				
Data				

The properties are chosen in such a way that they are apparently mutually exclusive. For the term "computer" there is no ambiguity as to which of the properties would apply. One would for instance get the corresponding row for the term "computer": 0 1 1 0, i.e., computer is not an abstraction, is a physical object, is hardware and not software. For the terms "system" and "program" it is not quite clear which of the properties apply. For instance program can be considered as an abstraction, if it is thought of as a mathematical algorithm, and on the other hand as a physical object if considered as a "written thing." Similarly system may be both hardware and software. Sometimes a term may be such that a property is neither applicable nor nonapplicable to it. It may be quite meaningless to ask whether the term "data" is software or hardware.

Of course if one would be able to choose such properties which are unambiguously mutually exclusive, then one would also be able to set up a hierarchical classification of terms from the term property matrix. For instance, in the previous example one could first get two mutually exclusive classes of terms: terms referring to abstract objects, and terms referring to physical objects. Then within these classes one would get the mutually exclusive classes of terms referring to hardware and software. The assumption of existence of mutually exclusive classes could be valid in subjects such as mathematics and the natural sciences, where the terms have unambiguous meaning (*37*). Because the terms in natural sciences are defined in terms of physically controlled experiments and the artificial language of mathematics there is much less ambiguity about them than for the terms of social sciences. Hence the construction of a thesaurus for engineering or atomic physics will have much less problems than that for sociology or political science. It is the experts in the field who determine the interrelationship of terms and in most cases there is a general agreement among them. The users of thesauri for such subjects will also be specialists, i.e., engineers or physicists who would associate terms in the way dictated by the purely technical development of the subject. The scope for a purely subjective approach is negligible. However it is quite untenable in social sciences (*35*).

It is important to point out that the term "property matrix" does not by itself lead to hierarchical classification of terms, but a classification is already assumed in the construction of the term property matrix.

• Need for a New Approach

Since a thesaurus is constructed for the information retrieval needs of a class of users, the interrelationship of terms established in the thesaurus should correspond to some "word map" which is common to the majority of this class of users. For subjects such as physics and engineering such a common "word map" can be built in consultation with a group of specialists in the field. In fact, this is a well known method, the so-called committee approach. Such a thesaurus will be used only by people having substantial knowledge about the subject. Hence one can reasonably assume that the thesaurus will effectively serve the information retrieval needs of this class of users. In fact, a thesaurus constructed on the basis of interrelationship of terms as viewed by the specialists, provided that in general there is sufficient agreement among them, could turn out to be more successful than the purely frequency-based automatic classification which lacks any logical or epistemological foundation. Of course the experiments with the statistical approach might show that these two approaches are not contradictory!

We would like to emphasize that a thesaurus should always be constructed with a class of users in mind. The construction of a thesaurus for a subject in the social sciences which will satisfy the information retrieval needs of users with many different ways of thinking appears to us to be as almost impossible. One tends to think that a thesaurus for sociology or political science should not only be useful for the specialists, but also for the common man. The dilemma of satisfying the incompatible needs of the specialists and the common man is faced by Viet (*38*). However he seems to assume that terms have some inner logical relationship independent of the relationships as viewed by the specialists. If it were assumed that concepts and relationships among concepts existed independently of the terms, as Soergel (*8*, p. 26) has assumed, then Viet's assumption would have been justified. Apart from the fact that such an assumption is epistemologically questionable, such an

Reproduced with permission of the copyright owner. Further reproduction prohibited without permission.

thesaurus is to be made is represented in the matrix form as follows:

	t_1	t_2	.	.	.	t_n
d_1	f_{11}	f_{12}	.	.	.	f_{1n}
d_2						
.						
.						
.						
d_m	f_{m1}	f_{m2}	.	.	.	f_{mn}

f_{ij} shows the frequency of the term t_j in the document d_i ($i = 1, 2, \ldots, m; j = 1, 2, \ldots, n$). Let us consider a simple example:

	t_1	t_2	t_3	t_4
d_1	4	1	0	5
d_2	3	0	1	4
d_3	0	4	6	1
d_4	1	3	4	0

In this example we find that t_1, t_4 are highly frequent in d_1 and d_2, whereas in d_3 and d_4 the frequency of both these terms is low. Similarly for t_2, t_3 we find that they are highly frequent in d_3, d_4 and are rare in d_1, d_2. Hence one concludes that t_1 is related to t_4, and t_2 is related to t_3. In this way classes of inter-related terms are formed so that terms belonging to the same class are characterised by the fact that they are almost equally frequent in a set of documents and almost equally rare in another set of documents.

As in the selection of key words the relationship among the terms is established purely statistically. The inference that these terms are also semantically related in some sense (like "beauty" is related to "love" in some sense) in a particular subject is only a hypothesis. However if it can be established, then one would even get a statistical theory of semantics.

The frequency approach has also been used to set up hierarchical relationship among terms (*34*). For a class of interrelated terms, those terms which have substantially higher frequencies are considered as categories and those with low frequencies are considered as subcategories. Hierarchical relationships based on frequency of terms might have been adequate for the information retrieval needs for a particular subject (*34*), but there does not seem to be enough ground to make it a general principle for thesaurus construction for all subjects. As we had pointed out earlier, setting up hierarchical relationships of terms is one of the most critical problems of library classifications (*35*). It cannot be solved simplistically by counting the frequency of terms. That a certain amount of human judgement is essential for setting up interrelationships and classification of terms is admitted also by Salton (*22*, p.57). The crucial problem is to determine how human judgement is to be used for establishing a classification as well as an interrelationship of terms.

Sparck-Jones (*36*) suggests a way of determining synonymity of terms by asking the decision maker whether a term can replace another term in a given context. If the synonymity of terms established in this way is to be of use for information retrieval, then the decision maker's answer should be representative for the whole class of users. But this is very difficult to assume, since his answer should not be personal one, and an element of subjectivity (and consequently ambiguity for information retrieval purposes) will inevitably creep in with many of the terms of social sciences. Consider for instance, the sentence "A person's behavior depends upon his education." If a decision maker would be asked whether in this sentence "behavior" can be replaced by "attitude," his answer would depend upon the fact whether he subscribes to behavioristic psychology or not. Such examples should caution us about the inherent problem of ambiguity of the terms of social sciences, and hence also about some of the fundamental difficulties of thesaurus construction.

"Question-answering systems" (*22*, p. 496-497) have been used to apply human judgement to the problem of classification of terms and determination of categories and subcategories. Like Dewey's rationalistic system, the categories and subcategories are given *a priori*, such as abstract, concrete, etc., and then questions are asked whether the terms have the properties expressed by the categories. In this way one gets a term property matrix:

	p_1	p_2	.	.	.	p_m
t_1	e_{11}	e_{12}	.	.	.	e_{1m}
t_2	.	.	.	.	.	.
t_n	e_{n1}	e_{n2}	.	.	.	e_{nm}

$e_{ij} = 1$ means that the property p_j is applicable to the term t_i ($i = 1, \ldots, n; j = 1, \ldots, m$). Similarly $e_{ij} = 0$ implies that p_j is not applicable to t_i.

Reproduced with permission of the copyright owner. Further reproduction prohibited without permission.

assumption would be unsound, even for the purposes of information storage and retrieval since this would force people to search for documents according to certain predetermined concepts and categories and not in their *own* way.

In subjects such as political science and sociology it might even be difficult to find agreement among the specialists as to the meaning and interrelationship of terms. However, to what extent there are variations in meaning can be explored by studying the association of terms used by a group of specialists. We shall briefly indicate how an analysis of the similarity and differences of such associations can be represented as graph theoretical models of different data structures. The focus of our problem is different from that of Quillian (*39*) but similar design principles can be employed to set up the "configuration of terms" for the meaning of words, where there is no intrinsic hierarchy among the terms, but the term whose meaning is sought becomes the primary focus, and other terms are seen only in relation to it:

> Most importantly, in such a model of semantic memory there is no predetermined hierarchy of superclasses and subclasses; every word is the patriarch of its own separate hierarchy when some search process starts with it. Similarly, every word lies at various places down, within the hierarchies of a great many other word concepts. Moreover, there are no word concepts as such that are "primitive." Everything is simply defined in terms of some ordered configuration of other things in the memory. (*39*, p. 415)

Quillian aims to simulate human memory by storing words in the memory of the computer with links corresponding to the relationship among the words as is given in a standard dictionary of the English language. For our problem of resolving the ambiguity of meaning, the standard dictionary is of no use. One can for instance consider some standard books or articles written by the specialists on a particular field. A questionnaire can then be designed to test the comprehension of some of the topics discussed by them. A multiple-choice questionnaire will prove quite convenient for processing the answers by a computer. Whatever the sample size of the population of readers, a certain degree of uniformity of the answers has to be there; otherwise, one has to admit that the degree of ambiguity is so great that communication between the author and the readers has been almost zero. The fact that even in social sciences a certain degree of effective communication is possible (in spite of the ambiquity of quite a few terms) is a guarantee that the answers to the questionnaire would help to approximate the "word maps" of different authors. The data structures corresponding to the word maps of different specialists can then be analyzed by a computer program which can then generate a data structure which is in a sense "a union" (more or less in the set theoretic sense of relations considered as classes of n-tuples) of the different data structures.

It is not necessary, as Reisner (*40-43*) has suggested, to make a "growing thesaurus" by extending the memory of the computer for different word relationships of each individual user if the particular relationship which a user thinks of has not already been stored. In fact, the idea of a growing thesaurus, which theoretically has to be infinite, is as impractical as the idea of a Turing machine with an infinitely long tape. Reisner seems to suggest that each individual has a unique word map. This might be true in some absolute sense. But much of the word map of an individual is to a large extent "programed" by the education he has received and the environment in which he has lived. In fact, really individualistic word relationships are of not much significance for the construction of a thesaurus, because there must be quite a few other people thinking in terms of such word relationships in order that some literature characterized by such a word relationship is available. In any case, if we assume that language is a sociological phenomenon, then the differences in word maps of different individuals should be to a large extent (statistically) explainable by differences in education and environment of different social groups within the same society. For the construction of a thesaurus for the social sciences it is not necessary to take into account the different shades of meaning with respect to each individual of the society but only according to certain groups, within which it is quite plausible to assume that a reasonable degree of uniformity exists. Such groups are, for example, the group of professors of social science, the group of students, the group of government officials, the group of women's liberationists, etc.

The ambiguity of terms in the social sciences should not discourage the construction of thesaurus for the social sciences. Naturally one should expect less exactness than one would from a thesaurus for engineering or nuclear physics. If one realizes that today the language of social sciences leaves much to be desired as far as clarity and precision are concerned, then one can understand that a thesaurus for the social sciences will have many more problems of ambiguity than the thesauri for the natural sciences.

• Conclusion

Recent work on thesaurus construction shows that it is generally assumed that for a given subject a hierarchical

Reproduced with permission of the copyright owner. Further reproduction prohibited without permission.

classification of terms exists. In the fully automatic approach one tries to establish such relationships, as also the selection of key words purely on the basis of frequency of occurence of terms in a given collection of documents. There is no logical or epistemological justification for this. However, this is not an unsound hypothesis. If it is verified by a large number of experiments then it could provide a basis for investigations into a statistical theory of semantics. In the semiautomatic approach the experts' answers to a set of questions give the clues for the classification. However, in the very framing of the question one has assumed *a priori* some hierarchical classification.

We have shown that it is necessary to distinguish the construction of thesauri for natural sciences from those for social sciences primarily because there is no general agreement today about the meaning of terms used in social sciences. However it might be possible to study the similarities and differences of interrelationships of words among different groups identified by similar interests. Such a study could adequately provide a basis for the construction of a thesaurus for social sciences.

References

1. Engineers Joint Council. 1967. *Thesaurus of Engineering and Scientific Terms.* New York: Engineers Joint Council; 1967.
2. Aitchison, J.; Day, P. eds. 1969. *Thesaurofacet: A Thesaurus and Faceted Classification for Engineering and Related Subjects.* Whetstone, Leicester: The English Electric Co. Ltd.; 1969.
3. Kinkade, R.G. 1974. *Thesaurus of Psychological Index Terms.* Washington, DC: American Psychological Association; 1974.
4. Vlet, J. 1971. *Thesaurus for Information Processing in Sociology.* Paris: Mouton & Co.; 1971.
5. American Political Science Association. 1975. *Political Science Thesaurus* (designed and developed at the University of Pittsburgh by the University Center for International Studies, the Knowledge Availability Laboratory, in conjunction with the Scientific Information Exchange Committee of the Association). Washington, DC: American Political Science Association; 1975.
6. Gilchrist, A. 1971. *The Thesaurus in Retrieval.* London: ASLIB; 1971.
7. Lancaster, F.W. 1972. *Vocabulary Control for Information Retrieval.* Washington, DC: Information Resources Press; 1972.
8. Soergel, D. 1974. *Indexing Languages and Thesauri: Construction and Maintenance.* Los Angeles, CA: Melville Publishing Company; 1974.
9. Wellisch, H.; Wilson, T.D. 1972. *Subject Retrieval in the Seventies.* College Park, MD: School of Library and Information Services, University of Maryland; 1972.
10. Jantsch, E. 1972. "Inter- and Transdisciplinary University: A Systems Approach to Education and Innovation." *Higher Education.* 1972; 1(1): 7-37.
11. Myrdal, G. 1975. "The Unity of the Social Sciences." *Human Organization.* 1975; 34(4):327-331.
12. Boulding, K. 1975. "Comment of The Unity on the Social Sciences." *Human Organization.* 1975; 34(4):332.
13. Ranganathan, S.R. 1959. *Elements of Library Classification.* London: Association of Assistant Librarians; 1959.
14. Ranganathan, S.R. 1971. *Colon Classification, 7th ed.* London: Asia Publishing House; 1971.
15. Gilchrist, A. (1971). *The Thesaurus in Retrieval.* London: ASLIB; 1971. 27.
16. Vickery, B.C. 1968. *Faceted Classification: A Guide to Construction and Use of Specialized Schemes,* 2nd ed. London: ASLIB; 1968.
17. Vickery, B.C. 1958. *Classification and Indexing in Science.* London: Butterworths; 1958.
18. Meincke, P. P.M.; Atherton, P. 1976. "Knowledge Space: A Conceptual Basis for the Organization of Knowledge." *Journal of American Society for Information Science.* 1976; 27(1): 18-24.
19. Vickery, B.C. 1963. Quoted and expanded in Bourne, C.P., *Methods of Information Handling.* New York: Wiley; 1963. 13-20.
20. Wall, E. 1969. "Vocabulary Building and Control Techniques." *American Documentation.* 1969; 20(2): 161-164.
21. Lesk, M.E. 1968. "Performance of Automatic Information Systems." *Information Storage and Retrieval.* 1968; 4:201-218.
22. Salton, G. 1968. *Automatic Information Organization and Retrieval.* New York: McGraw-Hill; 1968. 40-48.
23. Salton, G. 1972. "Experiments in Automatic Thesaurus Construction for Information Retrieval." In: *Information Processing 71, Proceedings of the IFIP Congress 1971, Vol. 1—Foundations and Systems.* Amsterdam: North Holland Publishing Company; 1972. 115-123.
24. Soergel, D. 1974. See (*8*) Chapter H, p. 554.
25. Sparck-Jones, K. 1971. *Automatic Keyword Classification and Information Retrieval.* London: Butterworths; 1971.
26. Wolff-Terroine, M.; Reinhart, D. 1971. "Computer-aided Automatic Generation of a Structured Documentary Language: 'A Preliminary Study." *Journal of Documentation.* 1971; 27(2):111-124.
27. Salton, G. 1968. "Automated Language Processing." In: *Annual Review of Information Science and Technology, Volume 3.* Chicago, IL: Encyclopaedia Britanica, Inc.; 1968.
28. Salton, G. 1975. *Dynamic Information and Library Processing.* Englewood Cliff, NJ: Prentice-Hall; 1975.
29. Long, J.M. et al. 1967. "Dictionary Buildup and Stability of Word Frequency in a Specialized Medical Area." *American Documentation.* 1967; 18(1):21-25.
30. Luhn, H.P. 1957. "A Statistical Approach to Mechanized Encoding and Searching of Literary Information." *IBM Journal of Research and Development.* 1957;1(4):309-317.
31. Luhn, H.P. 1961. "The Automatic Derivation of Information Retrieval Encodements for Machine Readable Texts." In: *Information Retrieval and Machine Translation, Part 2.* A. Kent, ed. New York: Interscience Publishers; 1961.
32. Doyle, L.B. 1965. "Is Automatic Classification a Reasonable Application of Statistical Analysis of Text?" *Journal of Association for Computing Machinery.* 1965;12(4):473-489.
33. Doyle, L.B. 1961. "Semantic Road Maps for Literature Searchers." *Journal of Association for Computing Machinery.* 1961;8(4):223-239.

Reproduced with permission of the copyright owner. Further reproduction prohibited without permission.

34. Doyle, L.B. 1965. "Expanding the Editing Function in Language Data Processing." *Communications of the ACM.* 1965;8(4):238-243.
35. Ghose, A.; Dhawle, A.S. 1975. "Inter- and Transdisciplinary Ordering System for Universe of Knowledge." Paper presented in the Third International Study Conference on Classification Research (Organized by FID/CR, in cooperation with FID/LD and FID/TM+OM) held at Bombay during January 6-11, 1975.
36. Sparck-Jones, K. 1965. "Experiments in Semantic Classification." *Mechanical Translation* 1965;8(3-4):97-112.
37. Ghose, A. 1973. *"Logic, Mathematics, Computer Science."* *International Logic Review.* 1973 June; 4(1):48-55.
38. Viet, J. 1971. *Thesaurus for Information Processing in Sociology.* Paris: Mouton & Company; 1971. 10.
39. Guilfilan, M.R. 1967. "Computers in Behavioral Science—Word Concepts: A Theory and Simulation of Some Basic Semantic Capabilities." *Behavioral Science.* 1967;12:410-430.
40. Reisner, P. 1965. "Construction of Authority Files." In: *Information Systems Compatibility.* Newman, S.M., Ed. London: Macmillan 1965. p. 55-67.
41. Reisner, P. 1965. "Semantic Diversity and a Growing Man-machine Thesaurus." In: *Some Problems in Information Science.* Kochen, M., Ed. New York: Scarecrow Press; 1965. 117-130.
42. Reisner, P. 1966. *Valuation of a "Growing" Thesaurus.* Yorktown Heights, NY: IBM Watson Research Center; 1966. 19p.
43. Borko, H. 1967. "Design and Information Systems and Services." In: *Annual Review of Information Science and Technology,* Vol. 2. New York: Interscience Publishers; 1967. 48.

Reproduced with permission of the copyright owner. Further reproduction prohibited without permission.

Thesaurus Construction And Maintenance: A Survival Kit
Batty, David
Database; Feb 1989; 12, 1; ABI/INFORM Global
pg. 13

THESAURUS CONSTRUCTION AND MAINTENANCE: A SURVIVAL KIT

by David Batty
CDB Enterprises Inc.

This article assumes that a well-developed thesaurus is based on a recognition of clusters of concepts that share common characteristics. Each cluster is organized as a hierarchy of concepts, and represented by natural language terms appropriate to the context in which the thesaurus will be used. Such an underlying structure aids and speeds the development of any kind of controlled index language, including a thesaurus.

The use of a thesaurus can make searching much easier and often more effective. Effective communication in any human context is improved by a common language that both parties understand, in this case, the use by both indexer and user of a prescribed vocabulary, with conceptual relationships clearly laid out, and scope notes and historical notes to encourage consistent application. Consult the Database Design column in the December, 1988 issue of *DATABASE* to read my "Introduction to the World of Thesaurus Building."

Planning and developing a thesaurus calls for a series of efforts. Among the first are finding out who the users are and what it is they need, the same points already considered in the course of planning the database that the thesaurus is to support. Then come these basic steps in thesaurus construction:

- Define the range and depth of the thesaurus
- Identify sources of the raw vocabulary
- Collect the raw vocabulary
- Organize the raw vocabulary into clusters of associated terms
- Shape the raw vocabulary into the word forms to be used in indexing the database
- Produce the thesaurus in appropriate display format

PLANNING THE THESAURUS

Who Are the Users?

How many users will there be, and will there be a wide range of skills? Will the users be knowledgeable in the subject area or will they need an extensive "entry" or "lead-in" vocabulary of many terms to lead them to the postable or preferred terms in the thesaurus? Are the users knowledgeable in using thesauri to explore databases? Can they find their way around the conceptual map of the thesaurus if they have to?

Will end-users be the most likely to be using terms from the thesaurus to search the database, or will someone such as a librarian or an information system person act as an intermediary? If end-users, what will be the level of information system awareness or expertise?

What do the Users Want?

Are the users likely to want a controlled index

►

Reproduced with permission of the copyright owner. Further reproduction prohibited without permission.

language, such as a thesaurus, at all? Some users prefer free text searching, in spite of the large retrieved sets it may yield, to what they see as the limitations of a controlled index language. ("I will not be dictated to by a...") Other users prefer a controlled index language to prompt them for search terms and to filter the information coming to them.

In manual systems, controlled index languages were essential. The size of any index attempting to accept all the terms used in the text of the documents would have been far too great. In computerized systems, text is accessible, but uncontrolled free text searching

> **...uncontrolled free text searching can produce undesirable results. . . a mix of free text searching and the use of a controlled index language has seemed a wise compromise.**

can produce undesirable results. For some time now, a mix of free text searching and the use of a controlled index language has seemed a wise compromise. Even the latest research in automatic indexing and retrieval systems based on artificial intelligence now assumes a domain language that is essentially an intellectually created controlled language.

Defining the Range and Depth of a Thesaurus

Determine the range of the thesaurus by defining the boundaries of the subject area that the thesaurus is to cover. Then decide on the depth of the thesaurus, in terms of the number of terms and their specificity.

Any special field of information can be divided into three broad areas:

- The *core area* of topics immediately pertinent to the main subject
- A *fringe area* of topics supportive of the main area, though not directly involved in it
- An *outside area* of ancillary subjects that may support the main area, but in which available subject expertise is likely to be at a minimum

Most of the terms with the greatest level of specificity are in the core area. Fewer terms, at decreasing levels of specificity, are in the fringe.

SUGGESTED READING, THESAURUS TERMINOLOGY & BUZZWORDS

MORE TO READ

AITCHISON, Jean
"Indexing languages: classification schemes and thesauri." In: *Handbook of special librarianship and information work*. d. L.J. Anthony. 5th ed. London: Aslib, 1982. Chapter 10.

Thesaurus construction: a practical manual, by Jean Aitchison and Alan Gilchrist. 2nd ed. London: Aslib, 1987. N.B.: without doubt the best book on thesaurus construction available.

AMERICAN NATIONAL STANDARDS INSTITUTE
ANSI Z39.19: Guideline for thesaurus structure, construction and use. New York: ANSI, 1980.

BATTY, C.D.
"Life begins at forty: the resurgence of faceted classification." In: *The Information Community: an alliance for progress*. Proceedings of the American Society for Information Science. Vol. 18. Washington, DC: ASIS, 1981. pp. 336, 340-342.

"Microcomputers in index language design and development." *Microcomputers for Information Management* 1, no .4 (December, 1984): pp. 303-312.

"The nature and development of multi-dimensional classification structures." In: *Information Communication and Technology Transfer*. Proceedings of the 43rd FID Congress, Montreal, 1986. The Hague: FID, 1988.

DUBOIS, C.P.R.
"The use of thesauri in online retrieval." *Journal of Information Science Principles and Practice* (Netherlands) 2 (March 1984): pp. 63-66.

FOSKETT, D.J.
The construction of a faceted classification for a special subject. London: Butterworth, 1959.

produced with permission of the copyright owner. Further reproduction prohibited without permission.

INTERNATIONAL ORGANIZATION FOR STANDARDIZATION
ISO2788: Guidelines for the establishment and development of monolingual thesauri. 2nd ed. Geneva: 150, 1986.

LANCASTER, F.W.
Vocabulary control for information retrieval. 2nd ed. Arlington, VA: Information Resources Press, 1986.

SOME THESAURUS TERMINOLOGY

POSTABLE TERM	Preferred term used to describe a topic
SCOPE NOTE	Definition of the use of a postable term
BT	Broader Term reference
NT	Narrow Term reference
RT	Related Term reference
HN	Historical Note to indicate changes in the use or form of a postable term
NOTATION	Formal symbolic code used to maintain a hierarchical arrangement of terms
TERM	Any word or phrase used to represent a concept
KEYWORD	Term encountered in source materials
DESCRIPTOR	Another word for POSTABLE TERM
NON-POSTABLE TERM	A non-preferred term entered in the thesaurus as a USE reference
CLUSTER	A group of terms which have basic semantic affinity
FACET	A group of terms which share a single characteristic

BUZZWORDS

Inevitably some terms clamoring for inclusion are suspect on the grounds of too great specificity or too short a life. Although such buzzwords should not be regarded as real thesaurus terms, neither should they be treated with disdain if users want them. Include them in a special list and put them in a special field in the database record so that those who want can retrieve information with those terms. Who now remembers the WESTERN LEG of the Alaskan natural gas pipeline, but ten years ago it was a hot item, a perfect buzzword.

—David Batty

Fortunately, by its very nature an outside area does not contain many terms, nor will they be very specific; here a good technical dictionary may substitute for the subject expert!

An example from the field of information science: INFORMATION RETRIEVAL, ONLINE DATABASE TECHNIQUES, and LEXICOGRAPHY lie in the core area, PUBLISHING and DATA PROCESSING are in the fringe area, and STATISTICS and EDUCATION are in the outside area.

Typically a thesaurus covers a special and easily identified area; sometimes a thesaurus must cover a broad area from a special point of view. For example a thesaurus on dental research has in its core TEETH, GUMS, PALATE, etc, their PROBLEMS, and the DENTAL PROCEDURES and MATERIALS to deal with the problems. The fringe and outside areas include other biomedical concerns, e.g., PAIN, ARTHRITIS (would you believe, the tooth is functionally a kind of joint?) or DRUGS.

On the other hand, a thesaurus for a newspaper database is apt to be more general, oriented perhaps towards GOVERNMENT, POLITICS, WORLD AFFAIRS, FINANCE, with strong elements of PEOPLE, SPORTS, INDUSTRY, SOCIAL SCIENCE, etc., with all topics only at a general level of subdivision, perhaps to be associated with whole text searching for greater specificity.

STEPS IN THESAURUS CONSTRUCTION

Identifying Sources of the Raw Vocabulary

If the thesaurus is for a specialized database of single-source literature, then a sample of that literature is the first obvious choice as a source of terminology. But if the still-specialized database may include other literature, as most do, then the best choice is a collection of several sources:

- A sample of the immediate literature, that which is actually used by the organization, is the first source
- A sample of the general literature of the subject, probably a conference proceedings or two, published by an organization central to the subject area, is another good source. Typically such publications have a hundred or more abstracts or papers. They also cover the field with a contemporary slant; the conference organizers will have balanced the program according to current preferences. You may want to supplement conference proceedings with ►

Reproduced with permission of the copyright owner. Further reproduction prohibited without permission.

current, principal journals that have abstracts.

• A sample of user questions, to analyze for concepts and directions, and to help to verify the terms used to request information. Many reference librarians or departments keep records of the questions that have been asked. Check these with other librarians who have worked with the users.

• Existing thesauri and indexes in the subject area, or related to it, are often valuable sources

This last point raises a fundamental question: is the thesaurus a revision of an existing language, or at least the sorting out of a list of terms already collected by the organization, and in use in some kind of (usually haphazard) indexing; or, are we to develop a thesaurus from scratch. The former is a very common situation. The points made already about the identification of sources still apply in such a case, though in a slightly different order of priorities.

Collecting and Recording the Raw Vocabulary

In the early stages of thesaurus development there should be no concern with the ultimate form of a term. For example, PAY, WAGES, SALARY, FEES, EMOLUMENTS are all words that represent PAYMENT FOR WORK DONE. We are concerned at this point with concepts; mere words may mislead us by their association sets. So, we select concepts expressed for the time being in the raw language of the source or in our own words if the term is contributed by us to fill a gap.

A typical abstract is 200-300 words long, and often yields ten or so key words or phrases principal to the subject of the document. For most special subject areas, several hundred abstracts will be sufficient to develop a basic, raw vocabulary. Documents without abstracts often have an executive summary, or an introduction that may be used in the same way, or they have at least chapter or section headings from the table of contents. Documents with no abstracts and with unbroken and lengthy text may be ignored in favor of another document!

Users' questions may not yield more than two or three keywords, since questions are inevitably simpler than the documents they need for the answers.

Existing thesauri and indexes yield terms as needed, within organized clusters, if you are lucky. Beware of existing thesauri that have a slightly different orientation; the organization of clusters (if any) may not be useful to you.

The traditional way to record raw vocabulary is on cards or slips of paper, designed as a form with boxes for the term under consideration, the source of the term, a scope note, references of different kinds and a notational code to represent the term and its position in a hierarchy of terms. At this early stage, only the term itself is entered and any other information relevant to it, like the source; we do not expect every detail or reference to be recorded now.

In this article the activities are described as if we are using paper slips. Although many people still find it convenient to use cards or paper slips, it is possible now to use microcomputers for the same purpose, either as storing, sorting and display devices, or in a more elaborate way. The use of microcomputers in thesaurus building and maintenance is discussed later in this article.

Organizing the Raw Terms into Clusters

Examine the whole set of raw terms for sets of associated terms, terms that have a basic semantic affinity, e.g., WOOD, PLASTIC, STEEL (materials, not the things that materials are made into), or DOOR, WINDOW, CEILING (parts of rooms, not rooms, nor the materials of which rooms are made), but not DOOR, WOOD (a functional, not a semantic, affinity).

Look for terms, record them, and then make piles of slips for THINGS (i.e., passive entities that something happens to), PROBLEMS, ACTIVITIES, AGENTS (i.e., active things that perform the activities), PLACES, etc. As these clusters of terms are built, be prepared to accept a "dump" pile, for terms that seem to belong nowhere. As the clustering proceeds, some piles will split, or scatter, and as they do, some terms from the dump pile will fit in. Sometimes more terms will join the dump pile. All this is natural. The sorting and organizing process seems endless and indeed can take up more than half of the thesaurus construction process, depending on the subject and the final thesaurus output.

Clusters of terms should be mutually exclusive, as far as possible, that is, no term in one cluster should appear in any other clusters without good reason. Sometimes they may: the term ASPIRIN may appear in both ORGANIC CHEMICALS, as a SALICYLIC ACID, and in DEPRESSANT DRUGS. ASPIRIN, then, has two BT (Broader Term) relationships: to the higher term in ORGANIC CHEMICALS and to the higher term in DRUGS. This is known as a polyhierarchy, something that lexicographers try to avoid if possible, but accept if inevitable.

Reproduced with permission of the copyright owner. Further reproduction prohibited without permission.

Clusters Within Clusters

In our attempt to arrive at well defined and easily scanned clusters of terms, we should recognize that sometimes a primary cluster may be a complex one. For example, if we have a cluster of terms for BUILDINGS, we will almost certainly have terms such as HOUSES, PREFABRICATED HOUSES, SINGLE-STORY DWELLINGS, that can fall into three separate subclusters: by the KIND of building, by the METHOD OF CONSTRUCTION, and by the HEIGHT, or NUMBER OF STORIES. The clue here is that we may easily imagine some combination like SINGLE-STORY, PREFABRICATED DWELLINGS, in which we have taken terms from different lists or at least from distinct parts of the same list.

Arrangement Within Clusters

Within the clusters we should find that we can arrange the terms in simple families, or hierarchies. These hierarchies should not be complex. Links to terms of a different kind (i.e., terms appearing in a different cluster) will be made later. Most relationships appearing within a cluster are parent/child or sibling relationships. Such hierarchies are easily discerned and made, and if later we wish to display the thesaurus with BT (Broader Term), NT (Narrow Term), and RT (Related Term) relationships of this kind, we shall find them already implicit in the hierarchies.

How Far Should We Go?

It is not too soon at this point to ask if we are not taking the effort of analysis too far. The answer lies mainly in the breadth and complexity of the vocabulary we are handling. Mrs. Fanny Beeton, in her famous English cookbook, often instructed, "Bake until done." Clearly we need to collect the raw vocabulary. If it is small, and if most of the terms are satisfactory in form, only a little tidying will be necessary to give us our thesaurus, and indexing can begin.

But if the raw vocabulary is too large to scan easily, then some clustering is necessary to allow us to begin the tidying process, which, by the way, must always be the last stage before we produce the controlled language in our chosen displays. Further, if the clusters themselves become too large, a subclustering process will probably be necessary.

Using Notation

Most people think of notation as a shelf ordering device in libraries. Essentially notation is simply an ordering device for any collection of objects, terms or concepts that have little intrinsic order, and, on which you have imposed an order, as we have just done in developing clusters, perhaps subclusters, and the hierarchies within them.

A study of notation can be fascinating to the point of addiction, but we should always remember that notation exists to preserve the order of our conceptual arrangement. If we have decided that ORANGES, LEMONS, LIMES and GRAPEFRUIT should be arranged *in that order*, we should not pervert it by using mnemonic notation in a different order: GR, LE, LI, OR.

For our purposes, (to quote W.C.B. Sayers, from his *Introduction to Library Classification*) "Notation should be easy to say, write, and remember." It should also be able to accommodate as many terms on one level as possible, and it should be able to accommodate new terms as the thesaurus grows. For this reason, a popular style of notation among lexicographers, especially of the English school, is an alphabetic notation which avoids the use of "I" "O" and possibly "L," depending on the use of upper and lower case letters, and the use of numeric characters. The use of punctuation marks, once popular, is now rare.

The notation should be applied evenly across an array of terms (siblings) in a family. If a term with the notation Md has six subordinate terms, Mdb, Mdf, Mdj, Mdn, Mdr and Mdv can be used. If there are more subordinate terms, close the gaps, and use Mdb, Mdd, Mdf, etc.

If an array of terms has more than 25 terms, consider breaking the terms into subgroups, and if that is too artificial, use a recognizably different notation to accommodate the large number of terms, e.g., 01-99 (which would always be used in two-digit groups to represent lists of terms up to one hundred in number).

Many thesauri today are used in machine-sensible environments, for development and maintenance and in indexing and retrieval. For this reason, notation is often also expressive or hierarchical, that is, a character of notation is added for every level of subdivision. This allows for efficient searching, but it can cause problems in thesaurus maintenance, where a new term may not be fitted in its proper place. In other words, the notation should be as full as needed to govern the order of the hierarchies, but ►

Reproduced with permission of the copyright owner. Further reproduction prohibited without permission.

should be as simple as possible for ease of comprehension.

Priority Order of Clusters

The innate order of notational characters implies an order of priority in the clusters that they are applied to. This is an unnecessary implication. If the resulting thesaurus is to be alphabetical, and the hierarchical display is simply to allow indexers and searchers to scan conceptually organized pieces of the vocabulary, then no order is needed, though within the clusters some order may suggest itself. Anyone with a passion for order (necessary or not) may wish to go back to the very early stage of defining the subject area in terms of its "core" and "fringe" sub-areas, and arrange the clusters in order from the central core onward and outward into the fringe.

Compound Terms or Precoordination

Everything that has been said so far has assumed that the vocabulary will be broken down into single-concept terms, and will so appear in the thesaurus. This is the generally accepted view, but there are times when compound terms are desirable, and almost inevitable.

Compound terms are usually either adjectival phrases (adjectives include nouns-as-adjectives) such as GUIDED MISSILES or SERIALS CONTROL, or prepositional phrases such as SAFETY IN MINES. They should be used as far as possible in their natural language form, as given here, and not in an inverted form such as MISSILES, GUIDED or MINES, SAFETY IN.

There are three general situations in which to include precoordinated terms in our vocabulary:

- The component terms are not used in any other sense, i.e., neither GUIDANCE nor MISSILES appears other than in GUIDED MISSILES (a rare situation)
- The phrase is very commonly used in the material to be indexed and by searchers
- The vocabulary is general enough so that the component terms in the phrase may be used in broader contexts, and the compound phrase is needed to clarify the particular meaning

The greater the number of precoordinated terms, when the component terms already exist, the larger the vocabulary, the more complex the reference structure, and consequently, the more laborious and expensive the maintenance effort. At the same time, a higher level of precoordination significantly increases the precision of retrieved sets of items, by eliminating "false drops," e.g., without precoordination, a search for EDUCATIONAL PROGRAMS IN COMPUTER LITERACY might also retrieve items on COMPUTER PROGRAMS and EDUCATION [for] LITERACY.

Refining the Raw Vocabulary

So far the vocabulary that represents the concepts in the index language has been left in its raw form. Now it must be refined into the form in which it will be used by the indexers and searchers.

First, *confirm the level of detail needed.* Any term below that level of specificity will be included as a USE reference to the nearest term above it in the generic hierarchy.

For example, THEORY OF SETS OF FINITE ORDER might appear as SETS OF FINITE ORDER in a thesaurus for mathematics, but in a general scientific thesaurus would more likely appear as a USE reference to SET THEORY or even to ALGEBRA.

Second, *translate the raw vocabulary into terms acceptable to users.* In a scientific thesaurus we would probably prefer MAGNESIUM SILICATE, with a reference from TALC, but in a popular thesaurus we might wish to reverse that reference. Phrases, as we have noted already, should be entered and used as they appear in natural language. In general the choice between singular and plural forms is determined by whether we can ask the question "how much" or "how many."

Third, *distinguish between the same word with different meanings or applications* in the thesaurus. For example, the word CELL used in BIOLOGY and in ELECTRICAL ENGINEERING is basically the same concept (and therefore represented by the same word) but with different applications. The word SIZE can mean MAGNITUDE and also a GLUTINOUS WASH used to prepare paper before painting or decorating. In such cases we can try to find acceptable synonyms so that our terms are unique, or we can add qualifiers to the term: CELL (BIOLOGY); CELL (ELECTRICAL ENGINEERING).

Fourth, *clarify the way in which a term should be used,* by adding a scope note. Be careful: a scope note is not a definition, though some definitions may become good scope notes. *A scope note is a definition of use.* A psychologists definition of ADOLESCENCE is not as useful to an indexer as the scope note information "Use for young people between the ages of 13 and 18."

Reproduced with permission of the copyright owner. Further reproduction prohibited without permission.

Making References

The development of a thesaurus as a set of hierarchies within mutually exclusive facets means that the generic/specific and sibling relationships are already indicated. How many of these will be included in the thesaurus display depends on the kind of display that we intend to use, but the potential is there for us to choose from. It remains to add the other kinds of references.

We have already looked at the conversion of potential BT/NT relationships to USE/USE FOR relationships, where the narrower term falls below a pre-determined level of specificity.

Sometimes there are cases where a term may have more than one BT (Broader Term). Typically this occurs where a term is included in a true generic/specific (THING/KIND OF THING) hierarchy, and also in WHOLE/PART hierarchy. Some purists claim that only the first of these can be truly a BT/NT relationship, and that the second is an RT (Related Term) relationship. But in practical terms, most users will accept, and even prefer, the BT/NT relationships in both cases. For example, ASPIRIN might appear in a facet of ORGANIC CHEMICALS under SALICYLIC ACIDS, and also in CENTRAL NERVOUS SYSTEM DEPRESSANTS. In such a case, the term ASPIRIN will show a thesaurus listing with two BT (Broader Term) relationships, and each generic term will show ASPIRIN as an NT (Narrow Term).

The most frequent addition to the reference structure is the RT (Related Term) relationship to a term outside the immediate hierarchy, e.g., BEES RT HONEY. This kind of reference cannot be derived automatically from the hierarchies. It must be recognized as a useful relationship and inserted deliberately by the lexicographer.

THESAURUS MAINTENANCE

Maintaining a thesaurus involves three kinds of activity and an overall management method.

Adding New Terms

New concepts represented in raw vocabulary will be encountered continually in indexing and searching. They may be contributed by the information staff or by end-users. They are handled in just the same way as in the initial construction of a thesaurus.

They are submitted as candidate terms, and considered for one of four possibilities:

- The candidate term is a new and necessary term that should be added to the thesaurus
- The candidate term is a synonym of a postable term already in the thesaurus
- The candidate term represents a useful concept but is too specific to be included
- The candidate term represents a concept not needed for the thesaurus

In the first case, the candidate term is included in the appropriate facet at the appropriate level of the hierarchy. The raw term is confirmed as, or changed to, the form of the term to be used in the vocabulary. Notation, scope note, and all references are added. In addition, all references are checked to ensure that reciprocal references appear at the terms referred to, thereby avoiding wild goose chases (a USE or SEE reference which has nothing in the place to which the user is referred!).

The second and third cases would result in a USE reference being made from the candidate term to the appropriate postable term. Obviously the last case is one for outright rejection, though a careful lexicographer might keep a list of those terms in case a change in the context or direction of the database resurrects such terms.

Deleting Thesaurus Terms

Indexing and searching may reveal terms that are rarely used, or that are misleading. Such terms may be deleted, or, with suitable editing, changed to USE references.

Amending Thesaurus Terms

Use of the thesaurus may reveal terms that are confusing or have typographical or spelling errors. Changes can range from a simple correction to the term as shown in the thesaurus, to a rewritten scope note, even to changes in the list of references.

Managing Thesaurus Maintenance

Suggestions for additions, deletions, or amendments should be channeled to a single office or person designated as the thesaurus editor or manager. Assisted by a small team of professional staff, subject experts, and perhaps end-users, the thesaurus editor should decide on a regular basis what additions or changes are to be made to the thesaurus. How often this occurs depends on the size of the thesaurus and the number of suggestions for change.

Users of the thesaurus should be informed as soon as possible of any changes. If the thesaurus ►

Reproduced with permission of the copyright owner. Further reproduction prohibited without permission.

is maintained in machine readable form, announcement of changes is comparatively easy. Significant changes in the scope or use of existing terms should not only be announced, they should also be recorded, either in a separate file, or preferably, in a special HISTORICAL NOTE in the thesaurus record itself, e.g.:

INFORMATION RESOURCE DICTIONARY SYSTEMS
HN Replaced Data Dictionaries, 1 Jan 82

Computer Maintenance of a Thesaurus

Even a quick look at available software for thesaurus management confirms everything that good indexers and lexicographers have known for many years: there is no real substitute for a *well thought out* controlled language with which to index and search databases. Computers can replace many of the tedious clerical tasks of thesaurus development, but since a computer cannot think creatively, there is an intellectual component in thesaurus development which this article has attempted to spell out.

The bad news first. There are no programs that can take that burden of thinking off your shoulders. There are programs that can assist in thesaurus management, which means that you must design the thesaurus, load it into the software and then use the software to manage new input, reciprocal references, assignment of scope notes, etc.

There are very few standalone software systems for thesaurus development and management for mainframe computers; most thesaurus management software is embedded in file- or database-management systems. At the microcomputer level there is nothing of a standalone character at the moment. Some microcomputer database management systems exist with a built-in thesaurus management capability. Two commercial packages that seem to be standalone come from England: STRIDE and TINTERM, but at the time of writing there are no details as to their availability. There is also a THESAURUS HANDLING SYSTEM written in dBASE II, produced by ICAITI, (Instituto Centroamericano de Investigacion y Tecnologia Industrial) supported originally by funds from the Organization of American States.

The microcomputer can be used in one of two ways: to list, sort, and display terms; or to add new terms, adjust the file and check new references and terms. The first of these two functions can be performed by any file or database management system that has commands for APPEND, BATCH-ADD, SORT, INDEX, and DISPLAY. The second requires programming to work within the thesaurus database.

Depending on the size of the vocabulary, the number of references, and whether the file management system handles fixed length fields or variable length fields, a special interest vocabulary of up to 5000 postable terms can be handled on a microcomputer with a 20MB hard disk. The lower threshold for the use of a machine readable thesaurus database is probably about 500 terms, and the upper level is at least 5000, which is certainly above the level that most lexicographers would expect to accommodate in a special subject.

ACKNOWLEDGMENT

My thanks to Betty Eddison of Inmagic for her time and suggestions for this article.

THE AUTHOR

DAVID BATTY has worked in, and run, every kind of library, except, regretfully, a bookmobile. He has been a reference librarian, cataloger, systems person, and teacher. He is now president of CDB Enterprises, and is a well-known consultant on information system design, indexing, thesauri and other index language construction.

Communications to the author should be addressed to David Batty, CDB Enterprises, 11608 Gilsan St., Silver Spring, MD 20902.

Reproduced with permission of the copyright owner. Further reproduction prohibited without permission.

ARTICLE IN PRESS

Available online at www.sciencedirect.com

SCIENCE DIRECT®

ELSEVIER

Physica A III (IIII) III–III

PHYSICA A

www.elsevier.com/locate/physa

Thesaurus as a complex network

Adriano de Jesus Holanda, Ivan Torres Pisa, Osame Kinouchi, Alexandre Souto Martinez*, Evandro Eduardo Seron Ruiz

Faculdade de Filosofia, Ciências e Letras de Ribeirão Preto (FFCLRP), Universidade de São Paulo (USP), Av. Bandeirantes, 3900, 14040-901 Ribeirão Preto, SP, Brazil

Abstract

A thesaurus is one, out of many, possible representations of term (or word) connectivity. The terms of a thesaurus are seen as the nodes and their relationship as the links of a directed graph. The directionality of the links retains all the thesaurus information and allows the measurement of several quantities. This has lead to a new term classification according to the characteristics of the nodes, for example, nodes with no links in, no links out, etc. Using an electronic available thesaurus we have obtained the incoming and outgoing link distributions. While the incoming link distribution follows a stretched exponential function, the lower bound for the outgoing link distribution has the same envelope of the scientific paper citation distribution proposed by Tsallis and Albuquerque (Eur. Phys. J. B 13 (2000) 777). However, a better fit is obtained by simpler function which is the solution of Ricatti's differential equation. We conjecture that this differential equation is the continuous limit of a stochastic growth model of the thesaurus network. We also propose a new manner to arrange a thesaurus using the "inversion method".
© 2004 Elsevier B.V. All rights reserved.

PACS: 05.90.+m; 02.50.−r

Keywords: Complex networks; Directed graphs; Thesaurus

Words are the building blocks to construct sentences and to transmit information. During the last decades much effort has been spent on the statistics of words. Concern has been centered in the similarities and differences among word distributions, which may be useful for application in automatic information retrieval and thesaurus construction. Zipf [1] has shown that word frequency obeys a power law if words are ranked from the most to the less frequent ones. Statistical linguistics, at its lowest level, can

* Corresponding author.
E-mail address: asmartinez@ffclrp.usp.br (A. Souto Martinez).

0378-4371/$ - see front matter © 2004 Elsevier B.V. All rights reserved.
doi:10.1016/j.physa.2004.06.025

be exemplified by the Zipf's exponent, which is very sensitive to the writer's instruction degree but much less sensitive to language (culture characteristics). Beyond word level, word connectivity has been treated in several manners. These treatments include entropic measures [2] and the construction of other quantities, such as the distribution of documents over the frequency of words [3]. Another interesting way to treat data is the latent semantic analysis (LSA) [4] which deals with word covariance in a corpus. LSA is a principal component analysis (PCA) technique, i.e., the covariance matrix is diagonalized and from the most important eigenvalues (around 300) the eigenvectors are considered to span a Euclidean vector space. A curious application of LSA is the automatic grading of high school texts [5]. However, LSA has been criticized as a poor approach for predicting semantic neighborhood [6]. Other studies have focused on a different approach. Words are tied to each other as links of a graph where the words are the nodes of it. Exhaustive studies over thesaurus [7,8] indicate that words are related among themselves as a small-world and scale-free network [9]. This means that words may be embedded in a low dimensional space but with a small fraction of long distance connections. The existence of the low-dimensional space has been suggested by the deterministic "tourist" walks [10,11] on the graph, which is an independent sampling procedure [12].

A *thesaurus* is a list of terms. A *term* can be a word, a composed word or even an expression. The list of related terms to a main entry term (head-word) provides alternatives for these entries. Following previous studies, we will consider terms as being "words" in a broad sense. As in a previous work [8], our study is based on unstructured thesaurus, the *Moby Thesaurus II* which is the largest[1] and most comprehensive free thesaurus data source in English available [13]. It has 30,260 (main) *entries*, also called *root words*[2] or *head-words*[3] and 73,046 words which are referred from the entries but they are not entries. They are called *non-root words*. These add up to 103,306 different words. Each root word points, in average, to 83 words.[4] The thesaurus-derived network is defined considering each term as a node. Connections are established from an entry to its related list of terms forming a *directed graph*. Classification of terms can be accomplished looking at the links (arcs), for instance: head-words (root words) are words with at least one emerging link ($k_{out} > 0$) and non-root words are words with no emerging links ($k_{out} = 0$). Apparently, there is a giant strong component (percolative cluster of directed links) which connects a large fraction of words [14,15]. We stress that the working thesaurus is a simple and unstructured *related term* thesaurus and we point out the existence of other thesauruses such as WordNet [16] and definition terms thesaurus (Roget's thesaurus) which may be modeled as a bipartite graph, but they are not considered here. If only *co-linked terms* (mutually referred terms) are considered, this structure forms a digraph and reduces

[1] The file has 24,271 KB.

[2] A *root word* should not be confused with a *root node* which is defined as one that has no incoming link.

[3] Some curiosities are: 877 words which are not referred from other entries, 16 words are entry words but point only to non-root words.

[4] From which, in average, 54 are root words and 29 are non-root words.

to the previous studied one, where its small-world and scale free structure has been pointed out [8]. In this case, the number of connections k of a node is called *degree of a node*. The node degree statistics shows an exponential behavior for small values of k and a power law behavior for large values of k [8].

The directed graph concepts permit us to classify sets of terms according to its links properties as follows:

Sink composed of the 73,046 terms with $k_{out} = 0$. For example: glucose, pass-word, all-around, grape juice, send word, put to, lap dog and afterbirth.

Source are the 30,260 terms with at least one outgoing link ($k_{out} > 0$), usually called main entries, entries, head-words or root words. The source can be divided into three categories.

Absolute source is related to 877 terms without incoming links $k_{in} = 0$. For example: rackets, grammatical, double quick, half moon, blinded.

Normal source are 29,333 terms that receive links and send links to other source and sink terms ($k_{out} > 0$ and $k_{in} > 0$). For example: ablation, analogy, call out, factitious, laid low, make a deal.

Bridge source they are the 16 terms without outgoing links to source terms ($k_{out}(\text{source}) = 0$), listed: androgyny, Christian sectarians, Congress, detector, electric meter, enzyme, Esperanto, etc., Geiger counter, ghetto dwellers, harp, in fun, lobotomy, penicillin, perversely and Senate.

These definitions are illustrated by the subgraphs 1–4 in Fig. 1.

It is interesting to observe that the outgoing link distribution could be fitted by the scientific-papers citation frequency curve (escort probability) [17]

$$f(k_{out}) = \frac{N_0}{[1 + (q-1)\lambda k_{out}]^{q/(q-1)}}, \tag{1}$$

with $N_0 = 654 \pm 7$, $\lambda = (1.66 \pm 0.01) \times 10^{-2}$ and $q = 0.95 \pm 0.01$ ($r^2 = 0.993$ and $\chi^2 = 77$), see Fig. 2. Also, the Tsallis' density distribution $f(k_{out}) = N_0[1 + (q-1)\lambda k_{out}]^{-1/(q-1)}$ has been employed with $N_0 = 617 \pm 14$, $\lambda = (1.5 \pm 0.1) \times 10^{-2}$ and $q = 0.95 \pm 0.03$

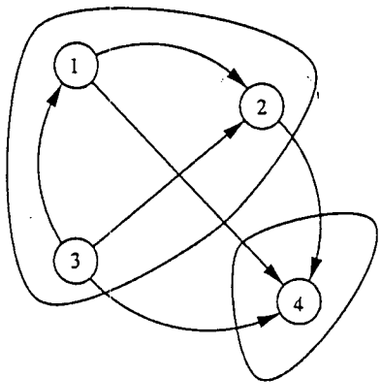


Fig. 1. Coarse grained view of the thesaurus as a directed graph. The region composed by subgraphs 1 to 3 is the *source* and subgraph 4 is referred as the *sink*. The source contains: the *normal source*, named as subgraph 1; *bridge source*, named as subgraph 2 and *absolute source*, here called subgraph 3.

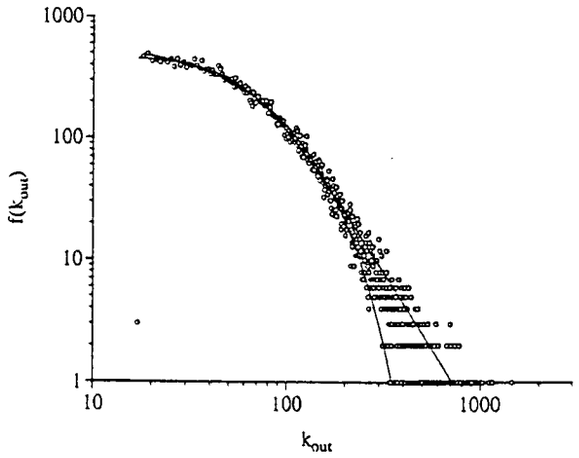


Fig. 2. The frequency of outgoing links k_{out} (root words) is well described by Eq. (2) which is the rightwards curve, in contrast with curve of Eq. (1). The point $[k_{out} = 17, f(k_{out}) = 3]$ has been excluded in both fitting procedures. These words are: for good, for keeps, and grin.

($r^2 = 0.95$ and $\chi^2 = 506$) which leads to a curve very close to the curve of Eq. (1) in Fig. 2. However, in both cases the limiting behavior for small k_{out} is exponential instead of the measured stretched exponential. So, we propose another fitting function with power law behavior for large k_{out} and stretched exponential behavior for small k_{out}

$$f(k_{out}) = \frac{N_0}{1 + \lambda k_{out}^{\kappa}}, \tag{2}$$

with $N_0 = 468 \pm 3$, $\lambda = (2.0 \pm 0.4) \times 10^{-5}$ and $\kappa = 2.55 \pm 0.03$ ($r^2 = 0.990$ and $\chi^2 = 99$), see Fig. 2. For $\lambda k_{out}^{\kappa} \ll 1$ this curve presents the same behavior of a stretched exponential [18]: $f(k_{out}) = N_0 \exp(-k_{out}/\bar{k}_{out})^{\kappa}$, which permits us to estimate the mean value of outgoing links equal to $\bar{k}_{out} = \lambda^{-1/\kappa} = 70 \pm 6$, which must be compared to 83 which is obtained from the thesaurus statistics. For $\lambda k_{out}^{\kappa} \gg 1$, the distribution of k_{out} is a power law: $f(k_{out}) = (N_0/\lambda) k_{out}^{-\kappa}$.

On the other hand, we show in Fig. 3 that the frequency of words with a given number of incoming links (k_{in}) is very well described by the stretched exponential curve

$$f(k_{in}) = N_0 \exp\left(-\frac{k_{in}}{\bar{k}_{in}}\right)^{\kappa}, \tag{3}$$

where we have found $N_0 = 12000 \pm 300$, $\bar{k}_{in} = 4.9 \pm 0.3$ and $\kappa = 0.52 \pm 0.01$, ($r^2 = 0.993$ and $\chi^2 = 4.58$). We shall stress that a fitting curve of the type of Eq. (2) also describe this data if λ is taken small enough. A simple approximation may be used as: $f(k_{in}) \propto \exp(-\sqrt{k_{in}})$. The low values of incoming links ($k_{in} < 10$) are dominated by non-root words while high values ($k_{in} > 100$) are dominated by root words, as seen in Fig. 3.

Although empirically $f(k_{in})$ and $f(k_{out})$ are apparently different, this may be due to a finite size database effect. This is suggested by a $k_{in} \times k_{out}$ plot (Fig. 4) where k_{in}

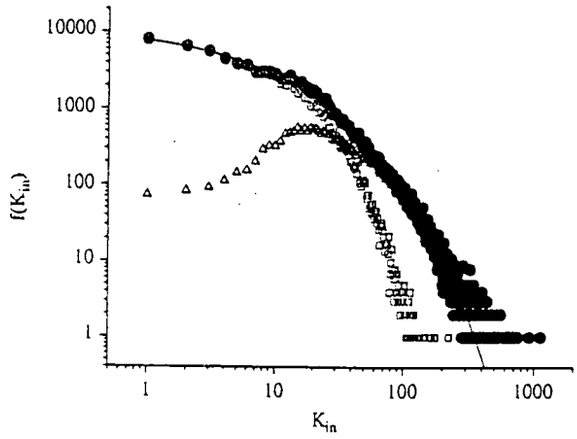


Fig. 3. Frequency of incoming links k_{in} for all words (•), root words (△) and non-root words (□). The curve for all words (•) is well described by a stretched exponential (line) expressed by Eq. (3) ($N_0 = 12000 \pm 300$, $\bar{k}_{in} = 4.9 \pm 0.3$ and $\kappa = 0.52 \pm 0.01$) which is dominated by non-root words for low k_{in} values ($k_{in} \leqslant 10$) and by root words for high k_{in} values ($k_{in} \geqslant 100$).

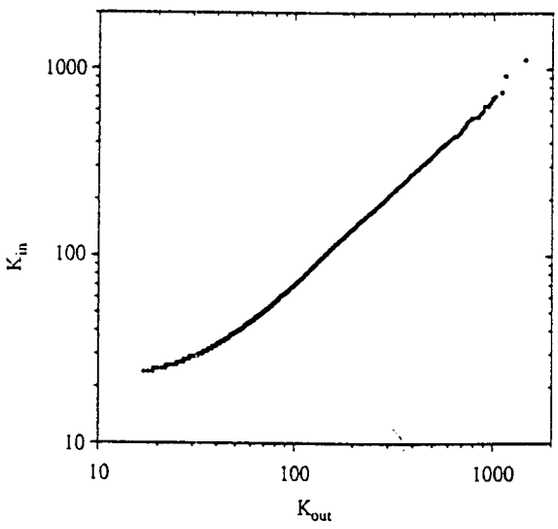


Fig. 4. The number of links k_{in} and k_{out} are ranked by decreasing values and plotted jointly to show the correlation.

and k_{out} are ranked by decreasing values and plotted jointly to show the correlation between them. From Fig. 4, it is clear that a linear correlation occurs for $k > 100$. A perfect thesaurus should have a symmetric property $k_{in} = k_{out}$.

As suggested by the above analysis, let Eq. (2) represent both the distribution of outgoing and incoming links. If one takes the variable to be continuous, it is not hard

to notice that Eq. (2) is the solution of the Ricatti's differential equation[5]

$$y'(x) = -\frac{\kappa\lambda x^{\kappa-1}}{N_0}\, y^2(x)\,. \tag{4}$$

This equation is known to represent contact processes such as the propagation of diseases [19]. Presently, we are searching for a microscopic network growth model that has the Ricatti's equation as a continuum limit.

A thesaurus is a attempt to synthesize terms and their relationships as natural as possible. Nevertheless, this trial is artificial and subjective. Our work of treating the thesaurus as a directed graph has provided new insights into its macro structure. From this graph theoretical approach the counting of k_{in} and k_{out} could lead to a novel proposition of term arrangement and term connectivities in it. The standard thesaurus classification is made according to word frequency in a corpus. But our approach suggested to rank the term related to a given root-word by its k_{in} ranking or any other connectivity index. For a finite thesaurus, where the number of root-words is smaller than the total number of terms, we suggest that it is always possible to construct a new closed thesaurus by inversion between the initial $k_{out} > 0$ root-words and $k_{in} > 0$ non-root words. The information in both cases are the same, but the latter leads to practical facilities, for instance, the fact that all words present in the thesaurus become root-words (the thesaurus becomes closed). We are submitting our closed version of Moby Thesaurus as a freeware database to the Gutenberg project [13].

The cumulative of most distributions considered here can be generalized by the BurrXII distribution: $f_q^{(\kappa)}(x) = (\kappa/\bar{x})(x/\bar{x})^{\kappa-1}[1 + ((q-1)/(2-q))(x/\bar{x})^{\kappa}]^{-1/(q-1)}$. This distribution can be obtained from the maximization of the non-extensive entropy with constraints on $\ln_q(x)$ and x^{κ} as noticed by Brouers et al. [20]. BurrXII distribution advantage is to describe a behavior typical of relaxation of many complex systems [21]. It has two power law behaviors, one for $x \ll \bar{x}$ and the other for $x \gg \bar{x}$ with a Weibull behavior in between ($x \sim \bar{x}$). Also, the BurrXII distribution reduces to Tsallis' distribution when $\kappa = 1$, $f_q^{(1)}(x) = \frac{1}{\bar{x}}[1 + ((q-1)/(2-q))(x/\bar{x})]^{-1/(q-1)}$ and to the Weibull distribution when $q = 1$, $f_1^{(\kappa)}(x) = (\kappa/\bar{x})(x/\bar{x})^{\kappa-1}\mathrm{e}^{-(x/\bar{x})^{\kappa}}$ (stretched exponential if $\kappa < 1$).

The authors are deeply grateful to Vera Lúcia Coelho Villar from the Instituto Antônio Houaiss de Lexicografia, Brazil, for the fruitful discussions. The authors thank stimulation discussions held with F. Brouers, M.G.V. Nunes, B.C.D. da Silva and C. Tsallis. This work has been partially funded by the Brazilian agencies: FAPESP, CAPES and CNPq.

References

[1] G.K. Zipf, Human Behavior and Principle of Least Effort, Addison-Wesley, Cambridge, MA, 1949.
[2] M.A. Montemurro, D.H. Zanette, Entropic analysis of the role of words in literary texts, cond-mat/0109218, 2001.

[5] Ricatti's differential equation is a particular instance of Bernoulli's differential equation [19].

[3] D. Volchenkov, P. Blanchard, S. Sharoff, Core lexicon and contagious words, cond-mat/0303454, 2003.
[4] T.K. Landauer, S.T. Dumais, A solution to Plato's problem: the latent semantic analysis theory of acquisition, induction, and representation of knowledge, Physcol. Rev. 104 (2) (1997) 211–240.
[5] W. Kintsch, The potential of latent semantic analysis for machine grading of clinical case summaries, J. Biomed. Inf. 104 (2002) 3–7.
[6] T.L. Griffiths, M. Steyvers, A probabilistic approach to semantic representation, http://www-psych.stanford.edu/~gruffydd/papers/semrep.pdf, last visited July 2004.
[7] M. Sigman, G.A. Gecchi, Global organization of the lexicon, cond-mat/0106509, 2001.
[8] A.E. Motter, A.P.S. de Moura, Y.-C. Lai, P. Dasgupta, Topology of the conceptual network of language, Phys. Rev. E 65 (2002) 065102(R).
[9] D.J. Watts, S.H. Strogatz, Collective dynamics of 'small world' networks, Nature (London) 393 (1998) 440–442.
[10] G.F. Lima, A.S. Martinez, O. Kinouchi, Deterministic walks in random media, Phys. Rev. Lett. 87 (1) (2001) 010603.
[11] H.E. Stanley, S.V. Buldyrev, Statistical physics—the salesman and the tourist, Nature (London) 413 (6854) (2001) 373–374.
[12] O. Kinouchi, A.S. Martinez, G.F. Lima, G.M. Lourenço, S. Risau-Gusman, Deterministic walks in random networks: an application to thesaurus graphs, Physica A 315 (3/4) (2002) 665–676.
[13] G. Ward, Moby Thesaurus II, Project Gutenberg Literary Archive Foundation, 2002, ftp://ibiblio.org/pub/docs/books/gutenberg/etext02/mthes10.zip.
[14] M.E.J. Newman, S.H. Strogatz, D.J. Watts, Random graphs with arbitrary degree distributions and their applications, Phys. Rev. E 64 (2) (2001) 026118.
[15] S.N. Dorogovtsev, J.F.F. Mendes, A.N. Samukhin, Giant strongly connected component of directed networks, Phys. Rev. E 64 (2) (2001) Art. No. 025101.
[16] Wordnet: A lexical database for the English language, http://www.cogsci.princeton.edu/~wn/, last visited July 2004.
[17] C. Tsallis, M.P. de Albuquerque, Are citations of scientific papers a case of nonextensivity? Eur. Phys. J. B 13 (2000) 777–780.
[18] J. Laherrère, D. Sornette, Stretched exponential distributions in nature and economy: 'fat tails' with characteristic scales, Eur. Phys. J. B 2 (1998) 525–539.
[19] W.E. Boyce, R.C. DiPrima, Elementary Differential Equation and Boundary Value Problem, 7th Edition, Wiley, New York, 2001.
[20] F. Brouers, O. Sotolongo Costa, K. Weron, Burr, Lévy, Tsallis, Physica A, these Proceedings, 2004, doi:10.1016/j.physa.2004.06.008.
[21] F. Brouers, private communication, November 2003.

Thesauri on the Web: current developments and trends

Ali Asghar Shiri and Crawford Revie

The authors

Ali Asghar Shiri is a PhD student and Crawford Revie is a Senior Lecturer, both in the Department of Information Science at the University of Strathclyde, Glasgow, UK.

Keywords

Information retrieval, Software, Internet, Indexes

Abstract

This article provides an overview of recent developments relating to the application of thesauri in information organisation and retrieval on the World Wide Web. It describes some recent thesaurus projects undertaken to facilitate resource description and discovery and access to wide-ranging information resources on the Internet. Types of thesauri available on the Web, thesauri integrated in databases and information retrieval systems, and multiple-thesaurus systems for cross-database searching are also discussed. Collective efforts and events in addressing the standardisation and novel applications of thesauri are briefly reviewed.

Electronic access

The current issue and full text archive of this journal is available at
http://www.emerald-library.com

Introduction

The rapid growth of networked information resources together with changes and innovations in the ways in which information is represented have necessitated a reassessment of the tools and techniques used for information management. World Wide Web technologies play a central role in refurbishing and redesigning information management tools.

Thesauri have played an important role in modern information storage and retrieval systems. While initial proposals to utilise thesauri focused on their ability to ensure consistent analysis of documents during input to information retrieval systems, they have increasingly become vital as aids to effective retrieval. Indeed, in the near future, it appears likely that thesauri will be used more during retrieval than at input (Milstead, 1998). The move to increasing use of thesauri as an aid to retrieval has expanded their functional span within information retrieval systems. As Aitchison *et al.* (1997) have noted, the role of the thesaurus is changing, but it is likely to remain an important retrieval tool.

This refocusing of the use of thesauri within information retrieval systems means that it is imperative that professionals take cognisance of the potential of thesauri as essential components of the largest information retrieval environment, namely the World Wide Web.

Why thesauri on the Web?

Although there are few operational information retrieval systems which have effectively incorporated thesauri as search and retrieval aids, we are witnessing an increased enthusiasm among thesaurus developers to make their thesauri available on the Web for potential applications. The reasons for this enthusiasm and the increasing availability of online thesauri are closely linked to key issues associated with the emergence of the World Wide Web, including:

- the colossal growth of information resources demanding their better subject identification;

Refereed article received 20 June 2000
Approved for publication 10 July 2000
This paper is based on a presentation delivered at the First Consultation on Agricultural Information Management (COAIM) conference, organised by FAO and held in Rome 5-7 June 2000.

Online Information Review
Volume 24 · Number 4 · 2000 · pp. 273–279
© MCB University Press · ISSN 1468-4527

Online Information Review
Volume 24 · Number 4 · 2000 · 273–279

- the migration of traditional information resources to the Web calls for more consistent subject approaches;
- an urgent need for resource description and discovery through reusing the existing information management tools such as controlled vocabularies;
- problems associated with quality of unstructured information retrieved from the Web;
- the need to provide users with knowledge structures such as thesauri for rapid and easy access to better organised information.

The emergence of vast amounts of information on the Web has created a new dimension in information retrieval research: Web-based information retrieval. As part of the attempt to apply information retrieval tools and techniques to manage information within this new medium in a more effective and efficient manner, the potential of thesauri to aid users' searching and browsing processes is being investigated.

Users, on the other hand, need good conceptual and semantic tools more than ever as they attempt to effectively organise the vast volumes of information available on the Web. The semantic structures provided by thesauri can play a part in both organising and retrieving Web information and knowledge resources.

Types of thesauri on the Web

Most thesaurus software producers are realising the potential of Web related programs such as HTML, JAVA and XML for thesaurus building and editing. Many of these producers have provided new versions of their traditional thesaurus management applications with more user-friendly Web interfaces and graphical displays together with navigation and browsing capabilities. These efforts in refurbishing traditional thesaurus applications have brought along thesaurus browsers and navigators.

Davies (1996) suggests that thesauri can be published over the Web in either static or dynamic form and the choice between the two methods of Web publication influences the format and organisation of the thesaurus.

A range of approaches has been used to publish thesauri on the Web, resulting in thesauri with varying levels of sophistication. Web-based thesauri can be categorised as follows in terms of their publishing format and structure:

- thesauri in simple static text formats (ASFA Thesaurus);
- thesauri in HTML format but still static, without effective use of hyperlinks (Inforterm);
- thesauri in dynamic HTML format with fully navigable hyperlinks (MeSH);
- thesauri with advanced visual and graphical interfaces (Plumb Design Visual Thesaurus);
- thesauri in XML format (Virtual HyperGlossary).

This diversity of tools and techniques has both helped and hindered thesauri as aids to searching and browsing the Web. The best examples have used graphical techniques to provide users with access to the rich cross-references and multiple levels of thesaural relationships possible using these approaches.

Web-based thesauri can also be divided into two general types in terms of their functionality and usability: standalone thesauri which are not a part of an information system (e.g. ASIS Thesaurus of Librarianship and Information Science) and those which are fully integrated into databases or information retrieval systems (e.g. the ERIC thesaurus, which is totally adopted and integrated within the ERIC database by SilverPlatter). References to a range of thesauri on the Web are provided at the "Controlled vocabularies resource guide" Website.

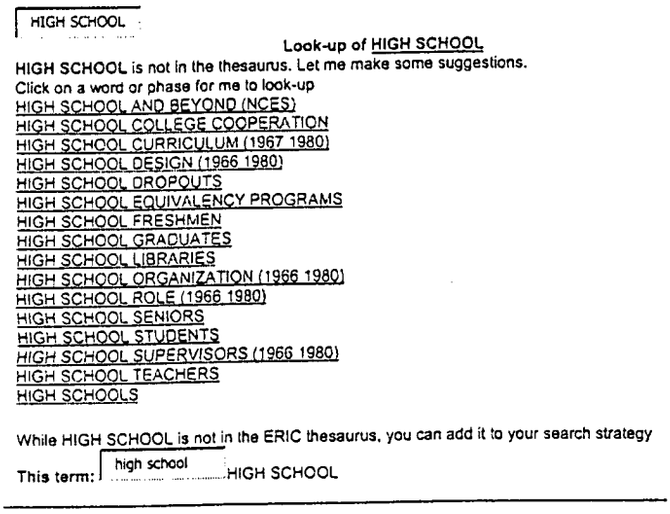
Thesauri as integral parts of databases

There are an increasing number of thesauri available on the Web covering a wide variety of subjects, formats, displays and languages. There are over 50 major thesauri covering around 30 subjects and the number is steadily increasing. The thesauri operating in connection with a database or an information retrieval system support different levels of user searching and browsing. Some major examples of thesauri that are fully integrated into databases are as follows.

The ERIC Thesaurus on the Web as implemented by the Assessment and Evaluation Clearinghouse provides sophisticated support in a sense that it attempts to provide the user with a suitable starting point and then to expand the query at different levels of thesaurus hierarchy in a user-friendly environment (see Figure 1). A useful feature of this implementation of the ERIC is that if the user enters a term, for

Online Information Review
Volume 24 · Number 4 · 2000 · 273–279

Figure 1 ERIC search wizard



example "high school", which is not in the thesaurus as a descriptor, the system informs the user and also provides several options of related terms to modify further browse and search entry.

Another example of a thesaurus which is explicitly used for both browsing and searching is the Humanities and Social Science Electronic Thesaurus (HASSET) developed and maintained by the University of Essex. HASSET was developed to allow subject access to the Data Archive, which is a specialist national resource containing the largest collection of accessible computer readable data in the social sciences and humanities in the UK. Using the HASSET subject thesaurus, users can extend their searches using different levels of hierarchy, e.g. broader or narrower terms. This makes the term selection process more flexible as users have access to both the thesaurus and the bibliographic database. HASSET also has different display formats with a variety of search and browse features.

The UK National Digital Archive Datasets (NDAD) which preserves and provides access to an emerging category of public record – computer datasets from UK government departments and agencies – is using a thesaurus whose construction is based on the UNESCO thesaurus. Searching the NDAD thesaurus will allow the user to find catalogues and administrative histories which relate to the subjects in the thesaurus.

The users can browse the thesaurus either hierarchically or alphabetically. Different thesaural relationships are nicely demonstrated and the user can easily choose among related, narrower or broader terms to be included in the search statement.

The MeSH thesaurus of the National Library of Medicine is a controlled vocabulary which consists of more than 19,000 main headings. In addition to these headings, there are 103,500 headings called Supplementary Concept Records. The MeSH browser may be used to find descriptors of interest and see these in relationship to other descriptors. This vocabulary aid is designed to help quickly locate descriptors of possible interest and to show the hierarchy in which these descriptors appear. The browser displays virtually complete MeSH records, including the scope notes, annotations, entry vocabulary, history notes, allowable qualifiers, etc. One of the most interesting features of the MeSH browser is its vast and extensive terms and vocabulary entries for different levels of searching. The user can browse through a MeSH tree structure or choose main headings, qualifiers and supplementary concepts. It provides a vast array of entry terms for users to choose from.

One of the most interesting examples of thesaurus integration with bibliographic databases is the work of Ovid Technologies. If the user enters a term, the term is automatically mapped to the subject headings of the database. By clicking on a subject heading the user can browse the thesaurus-related terms and choose the relevant terms. One of the notable search options of the Ovid products is the "focus" option through which users can limit their search to those documents in which the subject heading is considered the major point of the article.

Bowker-Saur's LISAnet is the Library and Information Science Abstracts database on the Web which allows the searchers to have easy access to its thesaurus for searching the database. Searchers can browse and search the terms in the thesaurus and through an readily accessible "paste" button, send the descriptor(s) to main search template for searching the database. It helps the searcher to easily identify and select alternative search terms and to add them for query reformulation purposes.

Multi-thesaurus searching and browsing systems

This enthusiasm for using modern Web technology to publish thesauri on the Web has resulted in a growing number of thesauri and the need to think of thesauri interoperability as well as the need for accessing and using

Online Information Review
Volume 24 · Number 4 · 2000 · 273–279

different thesauri for search and retrieval. As Fidel (1992) has showed in her research, better quality and availability as well as support for multi-database searching are likely to increase the use of controlled vocabularies. This is happening as we move toward using different thesauri for cross-database searching. The use of the Resource Description Framework (RDF), which was proposed by the World Wide Web Consortium, can also be considered one of the major breakthroughs in Internet-based application of thesauri and proving the ground for thesauri interoperability. The RDF namespace concept allows the controlled usage of distributed vocabulary systems. It also provides a syntax (XML) for exchanging controlled vocabulary data with other applications and services (Koch, 1999).

Currently, a number of projects are underway to enhance the level of access to multiple thesauri on the Web. One of the recent developments in this regard is the CERES (Californian environmental resources evaluation system) Thesaurus project for multiple environmental thesauri.

The purpose of this project is to construct an integrated controlled environmental vocabulary, and to develop the tools to use it for creation of metadata and construction of queries. Currently, the CERES program is involved in development of a key component, a controlled vocabulary database and user environment that will: provide pick list(s) for selection of keywords, synonyms, and related concepts for use in metadata and queries; provide a hierarchical organisation of information to serve as a browsing structure in information discovery; and allow simultaneous browsing and comparison of terms as presented in multiple standard thesauri in use nationally and internationally to index environmental information.

Knowledgecite Library (KCL) is a new online service which makes use of a multi-thesaurus tool for cross-database searching. If the user enters a term, knowledgecite will list thesaurus terms that match the user words or the phrase (e.g. cross-references, related, broader and narrower terms or scope notes) of all thesauri built into KCL, or those in the discipline (e.g. social sciences) that the user chose. Besides these thesaurus terms there are also cross-references to other thesauri (Jacso, 1999).

Multi-thesauri management system with Web interface is also another new development in using multiple thesauri. This project aims to provide a means for searching distributed databases of alternative medicine produced in various countries. The thesaurus management system performs at two levels, both with a Web interface: a search site open to anyone who wants to do cross-thesauri searching and browsing and a thesaurus maintenance site for editing the thesauri. At the moment, an alphabetical display, index of all terms, hierarchical display, and rotated display are the options for uploading, downloading, printing, and browsing. Once a hierarchical list or an alphabetical thesaurus is uploaded, other lists/displays can be generated automatically. Cross-thesauri searching for particular terms is performed based on exact and partial matching. Search terms are directly linked to the entries in a thesaurus. Our next step is to implement database searching.

GenThes or General Thesaurus Browser is another example of a facility for cross-references among thesauri which is able to handle several heterogeneous, multilingual thesauri. It is currently used by several environmental catalogue systems on the Web (Nikolai, 1999).

The Unified Medical Language System (UMLS) Metathesaurus is one of the largest thesauri in the area of medicine and related fields. It contains information about biomedical concepts and terms from many controlled vocabularies and classifications used in patient records, administrative health data, bibliographic and full-text databases and expert systems. The 2000 edition of the metathesaurus includes more than 730,000 concepts and 1.5 million concept names from over 50 different biomedical vocabularies, some in multiple languages. The metathesaurus is used in a wide range of applications including: linking between different clinical or biomedical vocabularies; information retrieval from databases with human assigned subject index terms and from free-text information sources; linking patient records to related information in bibliographic, full-text, or factual databases; natural language processing and automated indexing research; and structured data entry. The UMLS Metathesaurus is available to licensees via ftp, Web interface, and applications program interface (API) from the UMLS Knowledge Source Server. It is also available on CD-ROM by request.

The Consumer Health Terminology (CHT) Thesaurus provides consumers with online access to more than one million medical phrases and terms. The CHT Thesaurus is derived from Lexical's Metaphrase®

Online Information Review
Volume 24 · Number 4 · 2000 · 273–279

Thesaurus, which in turn is based upon the UMLS® Metathesaurus®, a compendium of controlled medical vocabularies maintained in part by Lexical under contract to the National Library of Medicine.

The CHT Thesaurus contains approximately one million medical definitions, phrases and terms, which have been inter-related over the past ten years. The CHT Thesaurus is especially useful to consumers because it includes more than 14,000 WellMed-developed consumer and lay terms that are mapped into these other vocabularies.

Using the CHT Thesaurus, search engines will be able to refine common search, providing users with the most appropriate information on the first try.

SNOMED, the Systematized Nomenclature of Medicine, is another comprehensive, multiaxial nomenclature created for the indexing of the entire medical record, including clinical findings, etiologies and interventions. As a reference terminology, SNOMED allows for consistent gathering and transmitting of detailed clinical information, retrieval of information for disease management or research and performance of outcome analysis for quality improvement.

Thesauri application in subject-based information gateways

Koch (2000) defines subject gateways as:

> Internet-based services which support systematic resource discovery. They provide links to resources (documents, objects or services), predominantly accessible via the Internet. Browsing access to the resources via a subject structure is an important feature.

Subject access through some kinds of knowledge structures like thesauri and classification systems is one of the significant features of good quality subject gateways. These quality-controlled subject gateways have established procedures for selection and content description of Web pages and also use thesauri for careful and consistent resource description.

Recently, several subject-based information gateways have been developed on the Web which use thesauri for indexing and retrieval of Web pages and Web sites. The following are some examples:

- Art, Design, Architecture and Media information gateway (Art and Architecture thesaurus);
- Engineering Electronic Library, Sweden (Engineering Information's EI thesaurus);
- Organising Medical Networked Information (Medical Subject Headings (MeSH) thesaurus);
- Social Science Information Gateway (HASSET thesaurus).

These subject gateways use thesauri to manually or automatically index Web pages and provide structured and more consistent subject access for browsing and searching the Web pages.

The Social Science Information Gateway (SOSIG), in its advanced search mode, provides a thesaurus search interface through which the user can enter a term and browse through and choose from the narrower, broader and related terms. The user can also send the query directly to SOSIG catalogue for searching the Web pages indexed by the term(s).

NKOS workshops: key discussions on thesaurus applications on the Web

The Networked Knowledge Organisation Systems (NKOS) workshops were initiated in 1997 within the ACM Digital Library Conference to address the issues of creating interactive knowledge organisation systems, including thesauri, over the Internet. The first workshop focused on using thesauri for searching and generating metadata. The use of thesauri in digital libraries, thesaurus-based metadata, subject vocabulary in distributed search environment and automated classification of networked information were among the main issues discussed in the workshop. This workshop aimed to gather information on research and products relating to the use of thesauri as metadata content tools.

The second workshop in 1998 was entitled "Application of terminology and classification tools for digital collection development and networked-based search". It focused on practical aspects of using thesauri in networked environments. Data models to support interactive use, intellectual property models, structural issues relating to terminology, including thesauri and XML, were among the topics addressed. In 1999, the NKOS workshop discussed standards for distributed thesauri on the Net. The application of thesauri to the Resource Description Framework (RDF), thesauri and metadata at Microsoft, and

Online Information Review
Volume 24 · Number 4 · 2000 · 273–279

multi-thesauri management systems with Web interfaces were among the issues for discussion.

This year, the NKOS 2000 workshop focused on ontology and related projects to develop a language for describing and exchanging ontologies as well as projects on constructing and maintaining special controlled vocabularies and terminologies for networked environments.

All the NKOS workshops have been significant events in defining the application of thesauri within the new information environment of the World Wide Web.

In search of a standard for electronic thesauri

The growing number of Web-based thesauri has stimulated discussion on the need to review existing standards for thesauri. The workshop on "Electronic Thesauri: Planning for a Standard" which was held in November 1999 and sponsored by the National Information Standards Organisation (NISO), American Society of Indexers (ASI), and the Association for Library Collections and Technical Service (ALCTS) was an attempt to address the issue of standards as they relate to thesauri. The workshop aimed to investigate the desirability and feasibility of a standard for electronic thesauri which:

- speaks to criteria and/or methods for generating thesauri by machine-aided or automatic means;
- shows semantic relationships among terms, as aids to text and information analysis and retrieval;
- supports a variety of electronic thesaurus displays;
- supports interoperability protocols, structures, and/or semantics applicable to thesauri.

Vocabulary mapping, management and interoperability were among the issues highlighted in the workshop. The need for flexible electronic thesaurus displays to ease and augment indexing and retrieval was also discussed. XML and RDF were put forward as widely supported formats to be useful for different browsing tools. It was pointed out that Web browsers are not thesaurus-aware and existing metadata formats make little use of thesauri. To address these issues it was suggested that special emphasis should be placed on the use of thesauri within metadata.

The workshop concluded that a standard for thesauri is needed and that this standard should provide for a broad group of controlled vocabularies such as classifications, taxonomies and subject headings. It also emphasised that the primary concern should be with interoperability rather than with construction or display.

Both this standards workshop and the NKOS events clearly demonstrate the need for and significance of thesauri as knowledge organisation tools. Moreover, one of the positive results of such joint efforts has been the emergence of a group of researchers and developers who are working toward creating and maintaining such knowledge organisation structures. This should provide good groundwork for future joint thesaurus projects and research initiatives.

Issues and areas of concern for the future

Due to the lack standards for publishing thesauri on the Web, a variety of formats, structures and features are currently in use. This causes problems concerning thesauri interoperability, reusability and shareability. There is an urgent need to examine semantic and syntactic tools, formats, and standards, used by Web-based thesaurus publishers and to find ways in which these aspects can be harmonised or integrated.

Many Web-based thesauri are not fully embedded as search and browse aids in databases, information retrieval systems and emerging Web search engines. These tools can be effectively utilised by search engines for more consistent and unified resource description and discovery.

Web-based thesauri can also be considered as tools for query formulation, refinement and expansion and can help users define their information needs more precisely and clearly. However, efforts are required to assess the extent to which these tools can contribute to more effective and reliable information retrieval within the context of the Web.

One of the major areas of concern relating to Web-based thesauri is the lack of evaluation studies which address their usability and functionality. At present, the growing number of thesaural interfaces with different searching and browsing features have not been subjected

to evaluative analysis. While traditional thesauri have been difficult to consult even for indexers, modern thesaurus browsers and navigators promise to become easy to use thanks to the Web-related technologies. Availability and accessibility of thesauri, particularly on the Web, has opened new doors for information users of all kinds and levels including scientists, researchers, and lexicographers. Therefore, it is of paramount importance to look into the variety of thesaurus consultation behaviours and to take account of users' preferences and perspectives in order to ensure their satisfaction. Research is also needed to shed light on how and for what purpose different types of people use thesauri and what value these tools can add to the information searching process in emerging information environments.

As noted earlier, some of the metadata formats on the Web are utilising thesauri for consistent subject assignment, yet it is not clear how much improvement thesaurus-based tagging will contribute to resource description and discovery. Furthermore, the feasibility of applying more structured and well-established thesauri onto different metadata formats is an issue for further investigation.

Concluding remarks

Electronic thesauri have been around for more than three decades and have greatly aided the process of information organisation. The advent of the Web together with recent developments in the application of thesauri as retrieval rather than organisation tools have brought about a new generation of thesauri. These tools are finding their way into a variety of Web-based information organisation and retrieval environments. Examples of their use include the application of thesauri in metadata, Web site/page indexing, Web accessible databases and Web search engines. Further research is needed to evaluate thesauri usability, functionality and effectiveness as distributed knowledge organisation and retrieval tools.

References

Aitchison, J., Gilchrist, A. and Bawden, D. (1997), *Thesaurus Construction and Use: A Practical Manual*, 3rd ed., Aslib, London.

Art, Design, Architecture and Media Information Gateway, http://adam.ac.uk/

Californian Environmental Resources Evaluation System (CERES) Thesaurus http://ceres.ca.gov/thesaurus/thesaurus_tool.html

Consumer Health Thesaurus http://www.cap.org/html/public/thesaurus.html

Controlled vocabularies resource guide http://www.fit.qut.edu.au/InfoSys/middle/cont_voc.html

Davies, R. (1996), "Publishing thesauri on the World Wide Web", *Advances in Classification Research Proceedings of the 7th ASIS SIG/CR Classification Research Workshop Held at the 59th ASIS Annual Meeting*, Baltimore, MD, October, pp. 44-54.

Engineering Electronic Library, Sweden http://eels.lub.lu.se

ERIC search wizard http://www.ericae.net/scripts/ewiz

Fidel, R. (1992), "Who needs controlled vocabularies?", *Special Libraries*, Vol. 83, Winter, pp. 1-9.

Humanities and Social Science Electronic Thesaurus (HASSET) http://dasun1.essex.ac.uk/services/zhasset.html

Jacso, P. (1999), "Savvy searchers do ask for direction", *Online & CD-ROM Review*, Vol. 23 No. 2, pp. 99-102.

Koch, T. (1999), "Automatic classification and content navigation support for Web services. DESIRE II Cooperates with OCLC", June 2000 www.oclc.org/oclc/research/publications/review98/koch_vizine-goetz/automatic.htm#Zthes

Koch, T. (2000), "Quality-controlled subject gateways: definitions, typologies, empirical overview", *Online Information Review*, Vol. 24 No. 1, pp. 24-34.

LISAnet (Bowker-Saur) http://www.lisanet.co.uk

Medical Subject Heading (MeSH) browser http://www.nlm.nih.gov/mesh/MBrowser.html

Milstead, J. (1998), "Thesauri in a full-text world", in Cochrane, P.A. and Johnson, E. (Eds), *Visualizing Subject Access for 21st Century Information Resources, Proceedings of the 1997 Annual Clinic on Library Applications of Data Processing*, Graduate School of Library and Information Science, University of Illinois, Urbana-Champaign, IL, pp. 28-38.

Multi-thesauri management system http://circe.slis.kent.edu/mzeng/thesaurihome.html

Networked Knowledge Organisation Systems (NKOS) workshops www.alexandria.ucsb.edu/~lhill/nkos

Nikolai, R. (1999), "GenThes: a general thesaurus browser for Web-based catalogue systems", www.computer.org/proceedings/meta/1999/papers/49/rnikolai.html

OMNI: Organising Medical Networked Information http://omni.ac.uk/

Ovid Technologies http://demo.ovid.com/demo/ovidWeb/ovidWeb.cgi

Social Science Information Gateway (SOSIG) http://www.sosig.ac.uk

Systematized Nomenclature of Medicine (SNOMED) http://ncvhs.hhs.gov/990517t5.htm

UK National Digital Archive Datasets (NDAD) Thesaurus http://ndad.ulcc.ac.uk/search/thesaurus.htm

Unified Medical Language System (UMLS) Metathesaurus http://www.nlm.nih.gov/pubs/factsheets/umlsmeta.html

Workshop of electronic thesauri: planning for a standard http://www.niso.org/thesau99.html