Overview and Framework for Data and Information Quality Research

STUART E. MADNICK and RICHARD Y. WANG
Massachusetts Institute of Technology
YANG W. LEE
Northeastern University
and
HONGWEI ZHU
Old Dominion University

Awareness of data and information quality issues has grown rapidly in light of the critical role played by the quality of information in our data-intensive, knowledge-based economy. Research in the past two decades has produced a large body of data quality knowledge and has expanded our ability to solve many data and information quality problems. In this article, we present an overview of the evolution and current landscape of data and information quality research. We introduce a framework to characterize the research along two dimensions: *topics* and *methods*. Representative papers are cited for purposes of illustrating the issues addressed and the methods used. We also identify and discuss challenges to be addressed in future research.

Categories and Subject Descriptors: A.1 [**Introductory and Survey**]; H.0 [**Information Systems**]: General

General Terms: Design, Experimentation, Management, Measurement, Reliability, Security, Theory, Verification

Additional Key Words and Phrases: Data quality, information quality, research topics, research methods

ACM Reference Format:
Madnick, S. E., Lee, Y. W., Wang, R. Y., and Zhu H. 2009. Overview and framework for data and information quality research. ACM J. Data Inform. Quality 1, 1, Article 2 (June 2009) 22 pages. DOI = 10.1145/1515693.1516680. http://doi.acm.org/10.1145.1515693.1516680.

Authors' addresses: S. E. Madnick, MIT Sloan School of Management, Cambridge, MA 02139; email: smadnick@mit.edu; Y. W. Lee, Department of Information, Operations and Analysis, College of Business Administration, Northeastern University, Boston, MA 02115; R. Y. Wang, MIT Information Quality Program, Cambridge, MA 02139; H. Zhu, College of Business and Public Administration, Old Dominion University, Norfolk, VA 23529.

1. INTRODUCTION

Organizations have increasingly invested in technology to collect, store, and process vast quantities of data. Even so, they often find themselves stymied in their efforts to translate this data into meaningful insights that they can use to improve business processes, make smart decisions, and create strategic advantages. Issues surrounding the quality of data and information that cause these difficulties range in nature from the technical (e.g., integration of data from disparate sources) to the nontechnical (e.g., lack of a cohesive strategy across an organization ensuring the right stakeholders have the right information in the right format at the right place and time).

Although there has been no consensus about the distinction between data quality and information quality, there is a tendency to use *data quality* to refer to technical issues and *information quality* to refer to nontechnical issues. In this article, we do not make this distinction and use the term *data quality* to refer to the full range of issues.

To address data quality concerns, researchers at MIT started investigating issues such as interdatabase instance identification [Wang and Madnick 1989] and data source tagging [Wang and Madnick 1990]. In 1992, the MIT Total Data Quality Management (TDQM) program was formally launched to underscore data quality as a research area [Madnick and Wang 1992]. Pioneering work at the TDQM program laid a foundation for data quality research and attracted a growing number of researchers to conduct cutting-edge research in this emerging field. The substantial output of this research community has been a primary driver for the creation of this *ACM Journal of Data and Information Quality (JDIQ)*.

In this article, we provide an overview of the current landscape of data quality research and discuss key challenges in the field today. We do not attempt to provide a comprehensive review of all, or even most, prior work in the field. Instead, for each topic and method we cite representative works to illustrate the range of issues addressed and methods used in data quality research. The cited works also serve as pointers for interested researchers to other relevant sources in the literature.

The rest of the article is organized as follows. In Section 2 we briefly review some of the pioneering work in data quality. In Section 3 we present a framework for characterizing data quality research. In Section 4 we describe various research topics in data quality research and cite a sample of works to exemplify issues addressed and the kinds of contributions the *JDIQ* will solicit. In Section 5 we review data quality research methods and show how they have been used to address a variety of data quality issues. Finally, in Section 6, we conclude the article with a brief discussion on some of the challenges that lie ahead in data quality research.

2. THE EVOLUTION OF TDQM DATA QUALITY RESEARCH

Early data quality research focused mainly on developing techniques for querying multiple data sources and building large data warehouses. The work of Wang and Madnick [1989] used a systematic approach to study related data quality concerns. This work identified and addressed entity resolution issues that arose when integrating information from multiple sources with overlapping records. These researchers explored ways to determine whether separate records actually corresponded to the same entity. This issue has become known by terms such as *record linkage*, *record deduplication*, and *record matching*.

Later, Wang and Madnick [1990] developed a *polygen* (*poly* for multiple, *gen* for source) model to consider the processing of data source tags in the query processor so it could answer data quality-related questions such as "Where is this data from?" and "Which intermediary data sources were used to arrive at this data?" Follow-up research included the development of a modeling method (known as the *Quality Entity Relationship* model) to systematically capture comprehensive data quality criteria as metadata at the conceptual database design phase [Wang et al. 1993; Storey and Wang 1998] and used an extended relational algebra to allow the query processor to process hierarchical data quality metadata [Wang et al. 1995a]. This stream of research has led to impacts on modern database research and design such as data provenance and data lineage [Buneman et al. 2001], and other extensions to relational algebra for data security and data privacy management. More importantly, these early research efforts motivated researchers to embark on the systematic inquiry of the whole spectrum of data quality issues, which in turn led to the inauguration of the MIT Total Data Quality Management (TDQM) program in the early 1990's.

Early research at the program developed the TDQM framework, which advocates continuous data quality improvement by following the cycles of *Define*, *Measure*, *Analyze*, and *Improve* [Madnick and Wang 1992]. The framework extends the Total Quality Management (TQM) framework for quality improvement in the manufacturing domain [Deming 1982; Juran and Goferey 1999] to the domain of data. A key insight was that although data is, in fact, a product (or by-product) *manufactured* by most organizations, it was not treated nor studied as such. Subsequent research developed theories, methods, and techniques for the four cycles of the TDQM framework.

Define. A major breakthrough was to define data quality from the consumer's point of view in terms of *fitness for use* and to identify dimensions of data quality according to that definition via a systematic multistage survey study [Wang and Strong 1996]. Prior to this research, data quality had been characterized by attributes identified via intuition and selected unsystematically by individual researchers. Key data quality dimensions were uncovered using a factor analysis on more than 100 data quality attributes identified by the survey study. These dimensions have been organized into four data quality categories: accessibility, contextual, representational, and intrinsic.

Measure. A comprehensive data quality assessment instrument was developed for use in research as well as in practice to measure data quality in organizations [Lee et al. 2002]. The instrument operationalizes each dimension into four to five measurable items, and appropriate functional forms are applied to these items to score each dimension [Pipino et al. 2002]. The instrument can be adapted to specific organizational needs.

Analyze. This step interprets measurement results. Gap analysis techniques [Lee et al. 2002] reveal perceptual differences between data dimensions and roles about the quality of data [Strong et al. 1997]. The three major roles are data collectors, data custodians, and data consumers [Lee and Strong 2004]. Analysis identifies the dimensions that most need improvement and root causes of data quality problems.

Improve. In this step, actions are taken either to change data values directly or, often more suitably, to change processes that produce the data. The latter approach is more effective, as discussed in Ballou et al. [1998] and Wang et al. [1998] where steps towards managing information as a product are provided. In addition, technologies mentioned earlier such as polygen [Wang and Madnick 1990] and Quality Entity Relationship model [Storey and Wang 1998] can be applied as part of the continuous improvement process.

In addition to developing and enriching the TDQM framework, the TDQM program made significant efforts towards solidifying the field of data quality, broadening the impact of research, and promoting university-industry-government collaborations via publications, seminars, training courses, and the annual International Conference on Information Quality (ICIQ) started in 1996. With the help of the TDQM program, the University of Arkansas at Little Rock has established the first-of-its-kind Master's and Ph.D. data quality degree programs in the U.S. to meet the increasing demand for well-trained data quality professionals and to prepare students for advanced data quality research [Lee et al. 2007a].

Today, data quality research is pursued by an ever-widening community of researchers across the globe. In addition to ICIQ, other professional organizations have organized focused workshops on various areas within the field of data quality (e.g., SIGMOD Workshop on Information Quality in Information Systems and CAiSE Workshop on Information Quality). On the industry side of the data quality field, major software vendors have begun to implement data quality technologies in their product and service offerings. In government, data quality has become an important component in many e-government and Enterprise Architecture (EA) initiatives [OMB 2007]. In the private sector, organizations have adopted variations on the TDQM methodology. An increasing number of companies have appointed a Chief Data Officer (CDO) or senior executives with responsibilities similar to the CDO to oversee data production processes and manage data improvement initiatives. Some groups have started to use the title Information Strategists to signify that data quality has critical and compelling applicability for an organization's strategies.

In the meantime, data quality research faces new challenges that arise from ever-changing business environments, regulatory requirements, increasing varieties of data forms/media, and Internet technologies that fundamentally impact how information is generated, stored, manipulated, and consumed. Data quality research that started two decades ago has entered a new era where a growing number of researchers actively enhance the understanding of data quality problems and develop solutions to emerging data quality issues.

3. A FRAMEWORK FOR CHARACTERIZING DATA QUALITY RESEARCH

An early framework for characterizing data quality research was presented in Wang et al. [1995a]. It was adapted from ISO9000 based on an analogy between physical products and data products. The framework consisted of seven elements that impact data quality: (1) management responsibilities; (2) operation and assurance costs; (3) research and development; (4) production; (5) distribution; (6) personnel management; and (7) legal function. Data quality research in 123 publications up to 1994 was analyzed using this framework. Although this framework was comprehensive, it lacked a set of intuitive terms for characterizing data quality research, and thus was not easy to use. Furthermore, the seven elements do not provide sufficient granularity for characterization purposes.

To help structure our overview of the landscape of data quality research, we have developed a framework that is easier to use. Taking a pragmatic approach, we developed this framework based on two principles. First, research methods known and used by researchers have evolved over time and continue to be used in different disciplinary areas. Some methods overlap with others, but the methodological nomenclature offers a cue for researchers in corresponding research areas. Second, the types of research topics continue to evolve. Instead of coining distinct or orthogonal categories, we selected and combined commonly known categories from various research communities to encourage multidisciplinary research methods. Thus the framework has two dimensions, *topics* and *methods*, and is derived from a simple idea: Any data quality research project addresses certain issues (i.e., topics) using certain research methods. For each dimension, we have chosen a small set of terms (i.e., keywords) that have intuitive meanings and should encompass all possible characteristics along the dimension. These keywords are listed in Table I and their detailed explanations are provided in the next two sections. These topic and method keywords also are used to categorize papers submitted for publication in the *ACM Journal of Data and Information Quality*.

For ease of use, we have chosen intuitive and commonly used keywords, such as *organizational change* and *data integration* for the topics dimension, and *case study* and *econometrics* for the methods dimension. We have grouped the topics into four major categories. For the research methods, which are listed in alphabetical order, we have included terms with varying levels of specificity. For example, *econometrics* is more specific than *quantitative* method. This framework gives users the flexibility to choose a preferred level of specificity in characterization. When using the framework to characterize a particular piece of research, the researcher would choose one or more keywords from each dimension. For example, the paper "AIMQ: A Methodology for Information Quality Assessment" [Lee et al. 2002] addresses the *measurement* and *assessment* topic and uses a particular *qualitative* method (i.e., survey questionnaire) along with a *quantitative* method (i.e., statistical analysis).

We can also view the framework as a two-dimensional matrix where each cell represents a topic-method combination. We can place a research paper in

Table I. Topics and Methods of Data Quality Research

Topics	Methods
1. Data quality impact	1. Action research
1.1 Application area (e.g., CRM, KM, SCM, ERP)	2. Artificial Intelligence
1.2 Performance, cost/benefit, operations	3. Case study
1.3 IT management	4. Data mining
1.4 Organizational change, processes	5. Design science
1.5 Strategy, policy	6. Econometrics
2. Database related technical solutions for data quality	7. Empirical
2.1 Data integration, data warehouse	8. Experimental
2.2 Enterprise architecture, conceptual modeling	9. Mathematical modeling
2.3 Entity resolution, record linkage, corporate householding	10. Qualitative
2.4 Monitoring, cleansing	11. Quantitative
2.5 Lineage, provenance, source tagging	12. Statistical analysis
2.6 Uncertainty (e.g., imprecise, fuzzy data)	13. System design, implementation
3. Data quality in the context of computer science and IT	14. Theory and formal proofs
3.1 Measurement, assessment	
3.2 Information systems	
3.3 Networks	
3.4 Privacy	
3.5 Protocols, standards	
3.6 Security	
4. Data quality in curation	
4.1 Curation - Standards and policies	
4.2 Curation - Technical solutions	

a particular cell according to the topic addressed and the method used. It is possible to place one paper in multiple cells if the paper addresses more than one issue and/or uses more than one method. A paper that uses a more specific method can also be placed in the cell that corresponds to a more general method. Obviously, some cells may be empty or sparsely populated, such as cells corresponding to certain combinations of technical topics (e.g., data integration) and social science methods (e.g., action research). Researchers are encouraged to consider employing more than one research method, including one or more quantitative methods with one or more qualitative methods.

In the next two sections, we use the framework to describe the landscape of data quality research. We also provide descriptions of keywords and illustrate their uses by citing relevant literature.

4. RESEARCH TOPICS

Data quality is a multidisciplinary field. Existing research results show that researchers are primarily operating in two major disciplines: Management Information Systems (MIS) and Computer Science (CS). We encourage researchers in other areas also to engage in data quality research and we encourage researchers in one field to borrow theoretical and methodological traditions from other disciplines as well. As a result of its multidisciplinary nature, data quality research covers a wide range of topics. In the following we provide a categorization scheme of data quality research topics. The scheme is broad enough to encompass topics addressed in existing research and those to be explored in future research. The scheme includes four major categories, each having a number of subcategories. A particular research activity can be categorized into multiple categories if it addresses multiple issues or multiple aspects of a single issue.

4.1 Data Quality Impact

Research in this area investigates impacts of data quality in organizations, develops methods to evaluate these impacts, and designs and tests mechanisms that maximize positive impacts and mitigate negative ones. There are five subcategories.

4.1.1 *Application Area (e.g., CRM, KM, SCM, ERP).* Research in this category investigates data quality issues related to specific application areas of information systems such as Customer Relationship Management (CRM), Knowledge Management (KM), Supply Chain Management (SCM), and Enterprise Resource Management (ERP). For example, Mikkelsen and Aasly [2005] reported that patient records often contain inaccurate attribute values. These inaccuracies make it difficult to find specific patient records. In another study, Xu et al. [2002] developed a framework for identifying data quality issues in implementing ERP systems.

4.1.2 *Performance, Cost/Benefit, Operations.* Research in this area investigates the impact of data quality on the performance of organizational units (including individuals), evaluates the costs and benefits of data quality initiatives, and assesses the impact of data quality on operations and decision making. As suggested by Redman [1998], poor data quality can jeopardize the effectiveness of an organization's tactics and strategies. Poor data quality can be a factor leading to serious problems [Fisher and Kingma 2001]. The impact of data quality and information about data quality on decision making has been investigated in several studies [Chengular-Smith et al. 1999; Fisher et al. 2003; Jung et al. 2005; Raghunathan 1999]. Preliminary research has assessed the impact of data quality on firm performance [Sheng and Mykytyn 2002]. Another study [Lee and Strong 2004] investigated whether a certain mode of knowledge, or *knowing-why*, affects work performance and whether knowledge held by different work roles matters for work performance. A recent study has shown evidence that the relationship between information quality and organizational outcomes is systematically measurable and the measurements of information quality can be used to predict organizational outcomes [Slone 2006]. Still more research is needed to assess the impact of data quality on entities as diverse as individual firms and the national economy.

4.1.3 *IT Management.* Research in this area investigates interactions between data quality and IT management, for example, IT investment, CIO leadership, and IT governance. The "fitness for use" view of data quality positions data quality initiatives as critical to an organization's use of IT in support of its operations and competitiveness. Organizations have begun to move from reactive to proactive ways of managing the quality of their data. We expect to see more empirical studies that gauge their effectiveness and uncover other effects of proactive data quality management.

4.1.4 *Organizational Change and Processes.* Ideally, data should be treated as a product, which is produced through a data manufacturing process. As suggested in prior research, data quality improvement often requires changes in processes and organizational behaviors. Research in this area investigates interactions between data quality and organizational processes and changes. For example, Lee [2004] investigate data quality improvement initiatives at a large manufacturing firm which iteratively adapted technical data integrity rules in response to changing business processes and requirements. A longitudinal study builds a model of data quality problem solving [Lee 2004]. The study analyzes data quality activities portrayed by practitioners' reflection-in-action at five organizations via a five-year action research study. The study finds that experienced practitioners solve data quality problems by reflecting on and explicating knowledge about contexts embedded in, or missing from, data. The study also specifies five critical data quality contexts: role, paradigm, time, goal, and place.

4.1.5 *Strategy and Policy.* Research in this area investigates strategies and policies for managing and improving data quality at various organizational and institutional levels. For example, Kerr [2006] studied strategies and policies adopted by the healthcare sector in New Zealand. The study shows that the adoption of a Data Quality Evaluation Framework and a national Data Quality Improvement Strategy provides clear direction for a holistic way of viewing data quality across the sector and within organizations as they develop innovations through locally devised strategies and data quality improvement programs. Data quality strategies and policies at a firm level are laid out in Lee et al. [2006].

4.2 Database-Related Technical Solutions for Data Quality

Research in this area develops database technologies for assessing, improving, and managing data quality. It also develops techniques for reasoning with data quality and for designing systems that can produce data of high quality. There are six subcategories.

4.2.1 *Data Integration, Data Warehouse.* Information systems within and between organizations are often highly distributed and heterogeneous. For analytical and decision-making purposes, there is a need to gather and integrate data from both internal and external sources (e.g., trading partners, data suppliers, the Internet). Integration can be enabled via a flexible query answering system that accesses multiple sources on-demand or via a data warehouse that pre-assembles data for known or anticipated uses. For example, Fan et al. [2001] provided ways to integrate numerical data. Data integration improves the usability of data by improving consistency, completeness, accessibility, and other dimensions of data quality. It is still an active research area after more than two decades of extensive study.

Goh et al. [1999] and Madnick and Zhu [2006] present a flexible query answering system, named COIN for *CO*ntext *IN*terchange, which employs knowledge representation, abductive reasoning coupled with constraint solving, and query optimization techniques. The system allows users to query data in multiple sources without worrying about most syntactic or semantic differences in those sources. It has been observed that many alleged "data quality" problems actually have been "data misinterpretation" problems. By understanding the contexts of both data sources and data consumers, COIN attempts to overcome data misinterpretation problems. It converts data, when necessary, to forms users prefer and know how to interpret.

Two other issues addressed by data integration research are entity resolution (sometimes known as record linkage or record deduplication, which are discussed later) and schema matching. Schema matching research [Rahm and Bernstein 2001; Doan and Halevy 2005] develops techniques to automatically or semiautomatically match data schemas. The results can be used for a query answering system to rewrite queries using one schema to query against other matched schemas. The results can also be used to construct a global schema [Batini et al. 1986] for a data warehouse.

A data warehouse is often built via *Extract, Transform, Load* (ETL) processes and provides tools to quickly interrogate data and obtain multidimensional views (e.g., sales by quarter, by product line, and by region). A framework for enhancing data quality in data warehouses is presented in Ballou and Tayi [1999]. The Data Warehouse Quality project [Jarke et al. 1999] has produced a set of modeling tools to describe and manage ETL processes to improve data quality [Vassiliadis et al. 2001]. In addition to various data cleansing tools designed specifically for ETL, a flexible query answering system such as COIN can be used as a transformation engine in ETL processes.

4.2.2 *Enterprise Architecture, Conceptual Modeling.* Enterprise Architecture (EA) [OMB 2007; Schekkerman 2004; Zachman 1987] is a framework for understanding the structure of IT elements and how IT is related to business and management processes. EA allows an organization to align its information systems with its business objectives. This alignment is often accomplished by documenting, visualizing, and analyzing relationships between systems and organizational needs. Enterprise architecture methods have been widely used. For example, federal agencies in the U.S. are required to adopt a set of Federal Enterprise Architecture (FEA) methods in IT operations, planning, and budgeting [OMB 2007]. Research in this area develops technologies to inventory, visualize, analyze, and optimize information systems and link their functionality to business needs.

Conceptual modeling allows the design of databases to meet a set of specific business requirements. The Entity-Relationship (ER) model [Chen 1976] and its extensions are the most prevalent data modeling techniques. One important extension is to add data quality characteristics to an ER model. As illustrated in Wang et al. [1993] and Storey and Wang [1998], this extension captures data quality requirements as metadata at the cell level. Furthermore, the querying system can be extended to allow for efficient processing of data quality metadata [Wang et al. 1995a]. There is ongoing research in this area to develop modeling extensions and query answering mechanisms to accommodate the need to manage data quality-related metadata such as quality metrics, privacy, security, and data lineage.

4.2.3 *Entity Resolution, Record Linkage, Corporate Householding.* An entity, such as a person or an organization, often has different representations in different systems, or even in a single system. Entity resolution [Wang and Madnick 1989; Talburt et al. 2005], also known as record linkage [Winkler 2006] and object identification [Tejada et al. 2001], provides techniques for identifying data records pertaining to the same entity. These techniques are often used to improve completeness, resolve inconsistencies, and eliminate redundancies during data integration processes.

A corporate entity is often composed of multiple subentities that have complex structures and intricate relationships. There are often differing views about the structures and relationships of the subentities of a corporate entity. For example, the answer to "What was the total revenue of IBM in 2008?" depends on the purpose of the question (e.g., credit risk assessment or regulatory filing). The purpose would determine if revenues from subsidiaries, divisions, and joint ventures should be included or excluded. This phenomenon sometimes is known as the corporate household problem and in certain cases can be modeled as an aggregation heterogeneity problem [Madnick and Zhu 2006]. More corporate household examples can be found in Madnick et al. [2005]. Actionable knowledge about organizations and their internal and external relationships is known as corporate household knowledge [Madnick et al. 2001]. Corporate householding research develops techniques for capturing, analyzing, understanding, defining, managing, and effectively using corporate household knowledge. Preliminary results of using context mediation for corporate householding management can be found in Madnick et al. [2004].

4.2.4 *Monitoring, Cleansing.* Certain data quality problems can be detected and corrected either online as data comes in or in batch processes performed periodically. Research in this area develops techniques for automating these tasks. For example, a technique for detecting duplicate records in large datasets is reported in Hernandez and Stolfo [1998]. The AJAX data cleansing framework has a declarative language for specifying data cleansing operations [Galahards et al. 2001]. This declarative approach allows for separation of logical expression and physical implementation of data transformation needed for data cleansing tasks. The framework has been adapted for data cleansing needs in biological databases [Herbert et al. 2004].

4.2.5 *Lineage, Provenance, Source Tagging.* Data lineage and data provenance information, such as knowledge about sources and processes used to derive data, is important when data consumers need to assess the quality of the data and make appropriate use of this data. Early research in this area [Wang and Madnick 1990] developed a data model that tags each data element with its source and provides a relational algebra for processing data source tags. A more general model was later developed [Buneman et al. 2001] which can be applied to relational databases as well as to hierarchical data such as XML. While much prior work focused on developing theories, an effort at Stanford University [Widom 2005] has developed a database management system to process data lineage information as well as uncertainties in data. A method of evaluating data believability using data provenance is developed in Prat and Madnick [2008].

4.2.6 *Uncertainty (e.g., Imprecise, Fuzzy Data).* From a probabilistic viewpoint, there is a certain degree of uncertainty in each data element, or conversely, an attribute can probabilistically have multiple values. Numeric values also have a precision. Research in this area develops techniques for storing, processing, and reasoning with such data [Dalvi and Suciu 2007]. For example, Benjelloun et al. [2006] present a novel extension to the relational model for joint processing of uncertainty and lineage information. While certain tasks require data with high precision and low uncertainty, other tasks can be performed with data that is less precise and more uncertain. Thus there is the need to effectively use data of differing levels of precision and uncertainty to meet a variety of application needs. Kaomea and Page [1997] present a system that dynamically selects different imagery data sources to produce information products tailored to different user constraints and preferences. In other cases, trade-offs need to be made between certainty or precision and other metrics of data quality. A mechanism of optimizing the accuracy-timeliness trade-off in information systems design is given in Ballou and Pazer [1995].

4.3 Data Quality in the Context of Computer Science and Information Technology

Research in this area develops technologies and methods (except for the specific database-related techniques covered before) to manage, ensure, and enhance data quality. There are six subcategories.

4.3.1 *Measurement, Assessment.* To manage data quality, an organization first needs to evaluate the quality of data in existing systems and processes. Given the complexity of information systems and information product manufacturing processes, there are many challenges in obtaining accurate and cost-effective assessments of data quality. Research in this area develops techniques for systematic measurement of data quality within an organization or in a particular application context. The measurement can be done periodically or continuously. Lee et al. [2002] present a data quality assessment and improvement methodology that consists of a questionnaire to measure data quality and gap analysis techniques to interpret the data quality measures. Useful functional forms used for processing the questionnaire results are discussed in Pipino et al. [2002].

Data quality can also be assessed using other methods. Here we give two examples. Pierce [2004] suggests the use of control matrices for data quality assessment. Data quality problems are listed in the columns of the matrix, quality checks and corrective processes form the rows, and each cell is used to document the effectiveness of the quality check in reducing the corresponding data quality problem. To improve the computation efficiency of data quality assessments in a relational database, Ballou et al. [2006] developed a sampling technique and a method of estimating the quality of query results based on a sample of the database.

4.3.2 *Information Systems.* In the broad field of information systems, data quality research identifies data quality issues in organizations, investigates practices that enhance or deteriorate data quality, and develops techniques and solutions for data quality management in an organizational setting. For example, taking a *product* view of information [Wang et al. 1998], Shankaranarayan et al. [2003] developed a modeling technique, called IPMap, to represent the manufacturing process of an information product. Using a similar modeling technique, Ballou et al. [1998] illustrated how to model an information product manufacturing system and presented a method for determining quality attributes of information within the system. Lee et al. [2007a] developed a context-embedded IPMap to explicitly represent various contexts of information collection, storage, and use. In a five-year longitudinal study of data quality activities in five organizations, Lee [2004] investigated how practitioners solved data quality problems by reflecting on and explicating knowledge about contexts embedded in, or missing from, data, and the contexts of data connected with otherwise separately managed data processes (i.e., collection, storage, and use).

4.3.3 *Networks.* There are a multitude of networks that connect various parts of a system and multiple systems. Networks can consist of physical communications networks, logic and semantic linkages between different systems, connections between systems and users, or even connections among users, such as social networks. Research into such networks can provide insights into how data is used and how the quality of data changes as data travels from node to node. These insights can be used to optimize network topology and develop tools for analyzing and managing networks. For example, O'Callaghan et al. [2002] proposed a single-pass algorithm for high-quality clustering of streaming data and provided the corresponding empirical evidence. Marco et al. [2003] investigated the transport capacity of a dense wireless sensor network and the compressibility of data.

4.3.4 *Protocols, Standards.* Data quality can be affected by protocols and standards. Research in this area develops protocols and standards to improve the quality of data exchanged among multiple organizations or within a single organization. Data standards improve data quality in dimensions such as consistency, interpretability, accuracy, etc. However, when data standards are too cumbersome, users may circumvent the standards and introduce data that deviate from these standards. Thus research in this area also needs to study how protocols and standards impact data quality and how organizations can promote user compliance. In addition, the quality of the protocols or standards is also subject to quality evaluation. For example, Bovee et al. [2002] evaluated the quality of the eXtensible Business Reporting Language (XBRL) standard to see if its vocabulary is comprehensive enough to support the needs of financial reporting.

4.3.5 *Privacy.* Certain systems contain private information about individuals, for example, customers, employees, patients. Access to such information needs to be managed to ensure only authorized users view such data and only for authorized purposes. Privacy regulations in different jurisdictions impose different requirements about how private data should be handled. Violating the intended privacy of data would represent a failure of data quality. Although there have been commercial tools for creating privacy rules and performing online auditing to comply with regulations, there are still many challenges in developing expressive rules and efficient rule enforcement mechanisms. Recent research also addresses privacy preservation issues that arise when certain data must be disclosed without other private information being inferred from the disclosed data. Such research has focused on developing algorithms to manipulate the data to prevent downstream users from inferring information that is supposed to be private [Li and Sarkar 2006; Xiao and Tao 2006].

Privacy concerns engender multiple requirements. For example, one aspect of privacy, called *autonomy*, is the right to be left alone. The "do not call" list in the U.S. is an example of legal protection for an autonomy requirement of privacy. As modes of communication with customers evolve, future research needs to develop effective solutions for describing the various privacy requirements and designing systems to meet these requirements. Further complicating privacy issues, some requirements such as those for data provenance can simultaneously increase quality while compromising privacy.

4.3.6 *Security.* Data security has received increasing attention. Research in this area develops solutions for secure information access, investigates factors that affect security, and develops metrics for assessing overall information security across and between organizations. A recent study [Ang et al. 2006] extends the definition of information security in three avenues: (1) *locale* (beyond the boundary of an enterprise to include partner organizations), (2) *role* (beyond the information custodians' view to include information consumers' and managers' views), and (3) *resource* (beyond technical dimensions to include managerial dimensions). This research attempts to develop an instrument for assessing information security based on this extended definition.

4.4 Data Quality in Curation

Digital curation is an emerging area of study originated in the fields of library and information science. It involves selecting, preserving, and managing digital information in ways that promote easy discovery and retrieval for both current and future uses of that information. Thus digital curation needs to consider current as well as future data quality issues. Consider the accessibility dimension of data quality: Data preserved on 8-inch and 5-inch floppy disks has become nearly inaccessible because it is difficult to find a computer that is equipped with a compatible floppy drive. There are other technical and nontechnical issues that need to be considered. For example, implicit contextual information that is known today (and often taken for granted) and necessary to interpret data may become unknown to future generations requiring explicit capture now to ensure future interpretability of curated data.

4.4.1 *Curation—Standards and Policies.* Research in this area develops standards and policies to improve data curation processes and strategies. A collection of curation related standards can be found at http://www.dcc.ac.uk/diffuse/, a site maintained by the Digital Curation Centre in the U.K.

4.4.2 *Curation—Technical Solutions.* In addition to database-related concerns, there are issues inherent in curation processes such as manually added annotations. Such manual practices provide challenges for data provenance requirements. Buneman et al. [2006] developed a technique to track provenance information as the user manually copies data from various sources into the curated database. The captured provenance information can be queried to trace the origins and processes involved in arriving at the curated data.

5. RESEARCH METHODS

Just as there is a plethora of research topics, there is a wide range of research methods suitable for data quality research. We identify 14 high-level categories of research methods.

5.1 Action Research

Action research is an empirical and interpretive method used by researchers and practitioners who collaboratively improve the practices of an organization and advance the theory of a certain discipline. It differs from consultancy in its aim to contribute to theory as well as practice. It also differs from the case study method in its objective to intervene, not simply to observe [Baskerville and Wood-Harper 1996]. An example of this research method can be found in Lee et al. [2004], which studied how a global manufacturing company improved data quality as it built a global data warehouse.

5.2 Artificial Intelligence

The field of artificial intelligence was established more than fifty years ago and has developed a set of methods that are useful for data quality research. For example, knowledge representation and automatic reasoning techniques can be used to enable semantic interoperability of heterogeneous systems. As demonstrated in Madnick and Zhu [2006], the use of such techniques can improve the interpretability and consistency dimensions of data quality. Agent technologies can be used to automate many tasks such as source selection, data conversion, predictive searches, and inputs that enhance system performance and user experience.

5.3 Case Study

The case study is an empirical method that uses a mix of quantitative and qualitative evidence to examine a phenomenon in its real-life context [Yin 2002]. The in-depth inquiry of a single instance or event can lead to a deeper understanding of *why* and *how* that event happened. Useful hypotheses can be generated and tested using case studies [Flyvbjerg 2006]. This method is widely used in data quality research. For example, Davidson et al. [2004] reported a longitudinal case study in a major hospital on how information product maps were developed and used to improve data quality.

5.4 Data Mining

Evolving out of machine learning of artificial intelligence and statistical learning of statistics, data mining is the science of extracting implicit, previously unknown, and potentially useful information from large datasets [Frawley et al. 1992]. The data mining approach can be used to address several data quality issues. For example, data anomaly (e.g., outlier) detection algorithms can be used for data quality monitoring, data cleansing, and intrusion detection [Dasu and Johnson 2003; Petrovskiy 2003; Batini and Scannapieco 2006]. Data mining has also been used in schema matching to find 1-to-1 matches [Doan et al. 2001] as well as complex matching relationships [He et al. 2004]. While many data mining algorithms are robust, special treatment is sometimes necessary when mining data with certain known data quality issues [Zhu et al. 2007].

5.5 Design Science

There is an increasing need for better design of information systems as many organizations have experienced failed IT projects and the adverse effects of bad data. A systematic study of design science has been called for in the information systems community. With an artifact-centric view of design science, Hevner et al. [2004] developed a framework and a set of guidelines for understanding, executing, and evaluating research in this emerging domain. As more artifacts such as Quality ER [Wang et al. 1993; Storey and Wang 1998] and IPMap [Shankaranarayan et al. 2003] are created to address specific issues in data quality management, it is important that they are evaluated using appropriate frameworks, such as the one suggested in Hevner et al. [2004].

5.6 Econometrics

A field in economics, econometrics develops and uses statistical methods to study and elucidate economic principles. A comprehensive economic theory for data quality has not been developed, but there is growing awareness of the cost of poor quality data [Øvretveit 2000] and a large body of relevant literature in the economics of R&D (Research and Development) [Dasgupta and Stiglitz 1980] and quality [De Vany and Saving 1983; Thatcher and Pingry 2004]. As we continue to accumulate empirical data, there will be econometric studies to advance economic theory and our overall understanding of data quality practices in organizations.

5.7 Empirical

The empirical method is a general term for any research method that draws conclusions from observable evidence. Examples include methods discussed earlier such as action research, case study, statistical analysis, and econometrics. Studies based on user surveys are also considered to be empirical studies. For example, surveys were used in Wang and Strong [1996] to identify data quality dimensions and the groupings of these dimensions. A survey with subsequent statistical analysis was also used in Lee and Strong [2004] to understand the relationship between modes of knowledge held by different information roles and data quality performance and in Slone [2006] to uncover the relationship between data quality and organizational outcomes.

5.8 Experimental

Experiments can be performed to study the behavior of natural systems (e.g., physics), humans and organizations (e.g., experimental psychology), or artifacts (e.g., performance evaluation of different algorithms). For example, Jung et al. [2005] used human subject experiments to examine the effects of contextual data quality and task complexity on decision performance. Klein and Rossin [1999] studied the effect of error rate and magnitude of error on predictive accuracy.

5.9 Mathematical Modeling

Mathematical models are often used to describe the behavior of systems. An example of this research method can be found in Ballou et al. [1998] where a mathematical model is used to describe how data quality dimensions such as timeliness and accuracy change within an information manufacturing system. System dynamics, a modeling technique originated from systems and control theory, has been used to model a variety of complex systems and processes such as software quality assurance and development [Abdel-Hamid 1988; Abdel-Hamid and Madnick 1990], which are closely related to data quality.

5.10 Qualitative

Qualitative research is a general term for a set of exploratory research methods used for understanding human behavior. Qualitative research methods suitable for data quality research include action research, case study, and ethnography [Myers 1997]. Examples of data quality research that used action research and case study have been discussed earlier. Ethnography is a research method where the researcher is immersed in the environment of the subjects being studied to collect data via direct observations and interviews. The method was used in Kerr [2006], which studied data quality practices in the health sector in New Zealand.

5.11 Quantitative

Quantitative research is a general term for a set of methods used for analyzing quantifiable properties and their relationships for certain phenomena. Econometrics and mathematical modeling are examples of quantitative methods suitable for data quality research. See the preceding discussions for comments and examples of these method types.

5.12 Statistical Analysis

Statistical analysis of data is widely used in data quality research. For example, factor analysis was used in Wang and Strong [1996] to identify data quality dimensions from survey data. Furthermore, statistics is the mathematical foundation of other quantitative methods such as data mining and econometrics.

5.13 System Design, Implementation

This research method draws upon design methodology in software engineering, database design and data modeling, and system architecture to design systems that realize particular data quality solutions. Using this method, trade-offs in the feature space can be evaluated systematically to optimize selected objectives. Researchers often use this method to design and implement proof-of-concept systems. The COIN system [Goh et al. 1999] was developed using this research method.

5.14 Theory and Formal Proofs

This method is widely used in theoretical computer science research such as developing new logic formalism and proving properties of computational complexity. The method is useful in theoretical data quality research. For example, Shankaranarayan et al. [2003] applied graph theory to prove certain properties of IPMap. Fagin et al. [2005] formalized the data exchange problem and developed the computational complexity theory for query answering in data exchange contexts.

6. CHALLENGES AND CONCLUSION

Data quality research has made significant progress in the past two decades. Since the initial work performed at the TDQM program (see web.mit.edu/tdqm) and later the IQ program (see mitiq.mit.edu) at MIT, a growing number of researchers from computer science, MIS, and other disciplines have formed a community that actively conducts data quality research. In this article, we introduced a framework for characterizing data quality research along the dimensions of topic and method. Using this framework, we provided an overview of the current landscape and literature of data quality research.

Looking ahead, we anticipate that data quality research will continue to grow and evolve. In addition to solving existing problems, the community will face new challenges arising from ever-changing technical and organizational environments. For example, most of the prior research has focused on the quality of structured data. In recent years, we have seen a growing amount of semistructured and unstructured data as well as the expansion of datasets to include image and voice. Research is needed to develop techniques for managing and improving the quality of data in these new forms. New ways of delivering information have also emerged. In addition to the traditional client-server architecture, a service-oriented architecture has been widely adopted as more information is now delivered over the Internet to traditional terminals as well as to mobile devices. As we evolve into a pervasive computing environment, user expectations and perceptions of data quality will also change. We feel that the "fitness for use" view of data quality has made some of the early findings extensible to certain issues in the new computing environment. Other issues are waiting to be addressed by future data quality research.

Much current research focuses on individuals and organizations. A broader perspective at a societal or group level can also be pursued, for example, research on data quality as it relates to people with disabilities or the elderly. As seen in other areas, researchers can collaborate across continents to uncover new insights into how data quality shapes global performance. We envision an evolving set of topics and methods to address new sets of research questions by new generations of researchers and practitioners.

REFERENCES

ABDEL-HAMID, T. K. 1988. The economics of software quality assurance: A simulation-based case study. *MIS Quart. 12*, 3, 395–411.

ABDEL-HAMID, T. K. AND MADNICK, S. E. 1990. *Dynamics of Software Project Management*. Prentice-Hall, Englewood Cliffs, NJ.

ANG, W. H., LEE, Y. W., MADNICK, S. E., MISTRESS, D., SIEGEL, M., STRONG, D. M., WANG, R. Y., AND YAO, C. 2006. House of security: Locale, roles, resources for ensuring information security. In *Proceedings of the 12th Americas Conference on Information Systems*.

BALLOU, D. P., CHENGALUR-SMITH, I. N., AND WANG, R. Y. 2006. Sample-Based quality estimation of query results in relational database environments. *IEEE Trans. Knowl. Data Eng. 18*, 5, 639–650.

BALLOU, D. AND PAZER, H. 1995. Designing information systems to optimize accuracy-timeliness trade-off. *Inf. Syst. Res. 6*, 1, 51–72.

BALLOU, D. AND TAYI, G. K. 1999. Enhancing data quality in data warehouse environments. *Commun. ACM 41*, 1, 73–78.

BALLOU, D., WANG, R. Y., PAZER, H., AND TAYI, G. K. 1998. Modeling information manufacturing systems to determine information product quality. *Manag. Sci. 44*, 4, 462–484.

BASKERVILLE, R. AND WOOD-HARPER, A. T. 1996. A critical perspective on action research as a method for information systems research. *J. Inf. Technol. 11*, 235–246.

BATINI, C., LENZERINI, M., AND NAVATHE, S. 1986. A comparative analysis of methodologies for database schema integration. *ACM Comput. Surv. 18*, 4, 323–364.

BATINI, C. AND SCANNAPIECO, M. 2006. *Data Quality: Concepts, Methodologies, and Techniques*. Springer Verlag.

BENJELLOUN, O., DAS SARMA, A., HALEVY, A., AND WIDOM, J. 2006. ULDBs: Databases with uncertainty and lineage. In *Proceedings of the 32nd VLDB Conference*, 935–964.

BOVEE, M., ETTREDGE, M. L., SRIVASTAVA, R. P., AND VASARHELYI, M. A. 2002. Does the year 2000 XBRL taxonomy accommodate current business financial-reporting practice? *J. Inf. Syst. 16*, 2, 165–182.

BUNEMAN, P., CHAPMAN, A., AND CHENEY, J. 2006. Provenance management in curated databases. In *Proceedings of ACM SIGMOD International Conference on Management of Data*, 539–550.

BUNEMAN, P., KHANNA, S., AND TAN, W. C. 2001. Why and where: A characterization of data provenance. In *International Conference on Database Theory*, J. Van den Bussche and V. Vianu, Eds. Lecture Notes in Computer Science, vol. 1973. Springer, 316–330.

CHEN, P. P. 1976. The entity-relationship model: Toward a unified view of data. *ACM Trans. Database Syst. 1*, 1, 1–36.

CHENGULAR-SMITH, I., BALLOU, D. P., AND PAZER, H. L. 1999. The impact of data quality information on decision making: An exploratory analysis. *IEEE Trans. Knowl. Data Eng. 11*, 6, 853–865.

DALVI, N. AND SUCIU, D. 2007. Management of probabilistic data: Foundations and challenges. In *Proceedings of the ACM Symposium on Principles of Database Systems (PODS)*, 1–12.

DASGUPTA, P. AND STIGLITZ, J. 1980. Uncertainty, industrial structure, and the speed of R&D. *The Bell J. Econom. 11*, 1, 1–28.

DASU, T. AND JOHNSON, T. 2003. *Exploratory Data Minding and Data Cleaning*. John Wiley & Sons, Hoboken, NJ.

DAVIDSON, B., LEE, Y. W., AND WANG, R. 2004. Developing data production maps: Meeting patient discharge data 1.

DE VANY, S. AND SAVING, T. 1983. The economics of quality. *The J. Political Econ. 91*, 6, 979–1000.

DEMING, W. E. 1982. *Out of the Crisis*. MIT Press, Cambridge, MA.

DOAN, A., DOMINGOS, P., AND HALEVY, A. 2001. Reconciling schemas of disparate data sources: A machine learning approach. In *Proceedings of the ACM SIGMOD Conference*, 509–520.

DOAN, A. AND HALEVY, A. Y. 2005. Semantic-Integration research in the database community: A brief survey. *AI Mag. 26*, 1, 83–94.

FAGIN, R., KOLAITIS, P. G., MILLER, R., AND POPA, L. 2005. Data exchange: Semantics and query answering. *Theoretical Comput. Sci. 336*, 1, 89–124.

FAN, W., LU, H., MADNICK, S. E., AND CHEUNG, D. W. 2001. Discovering and reconciling data value conflicts for numerical data integration. *Inf. Syst. 26*, 8, 635–656.

FISHER, C., CHENGULAR-SMITH, I., AND BALLOU, D. 2003. The impact of experience and time on the use of data quality information in decision making. *Inf. Syst. Res. 14*, 2, 170–188.

FISHER, C. AND KINGMA, B. 2001. Criticality of data quality as exemplified in two disasters. *Inf. Manag. 39*, 109–116.

FLYVBJERG, B. 2006. Five misunderstandings about case study research. *Qualitative Inquiry 12*, 2, 219–245.

FRAWLEY, W. J., PIATEKSKY-SHAPIRO, G., AND MATHEU S, C. J. 1992. Knowledge discovery in databases: An overview. *AI Mag. 13*, 3, 57–70.

GALAHARDS, H., FLORESCU, D., SHASHA, D., SIMON, E., AND SAITA, C. A. 2001. Declarative data cleaning: Language, model and algorithms. In *Proceedings of the 27th VLDB Conference*, 371–380.

GOH, C. H., BRESSAN, S., MADNICK, S. E., AND SIEGEL, M. D. 1999. Context interchange: New features and formalisms for the intelligent integration of information. *ACM Trans. Inf. Syst. 17*, 3, 270–293

HE, B., CHANG, K. C. C., AND HAN, J. 2004. Mining complex matchings across Web query interfaces. In *Proceedings of the 9th ACM SIGMOD Workshop on Research Issues in Data Mining and Knowledge Discovery*, 3–10.

HERBERT, K. G., GEHANI, N. H., PIEL, W. H., WANG, J. T. L., AND WU, C. H. 2004. BIO-AJAX: An extensible framework for biological data cleaning. *SIGMOD Rec. 33*, 2, 51–57.

HERNANDEZ, M. AND STOLFO. 1998. Real-World data is dirty: Data cleansing and the merge/purge problem. *J. Data Mining Knowl. Discov. 2*, 1, 9–37.

HEVNER, A. T., MARCH, S. T., PARK, J., AND RAM, S. 2004. Design science in information systems research. *MIS Quart. 28*, 1, 75–105.

JARKE, M., JEUSFELD, M. A., QUIX, C., AND VASSILIADIS, P. 1999. Architecture and quality in data warehouse: An extended repository approach. *Inf. Syst. 24*, 3, 229–253.

JUNG, W., OLFMAN, L., RYAN, T., AND PARK, Y. 2005. An experimental study of the effects of contextual data quality and task complexity on decision performance. In *Proceedings of the IEEE International Conference on Information Reuse and Integration*, 149–154.

JURAN, J. AND GOFEREY, A. B. 1999. *Juran's Quality Handbook*. 5th ed. McGraw-Hill, New York.

KAOMEA, P. AND PAGE, W. 1997. A flexible information manufacturing system for the generation of tailored information products. *Decision Support Syst. 20*, 4, 345–355.

KERR, K. 2006. The institutionalization of data quality in the New Zealand health sector. Ph.D. dissertation, The University of Auckland, New Zealand.

KLEIN, B. D. AND ROSSIN, D. F. 1999. Data quality in neural network models: Effect of error rate and magnitude of error on predictive accuracy. *Omega 27*, 5, 569–582.

LEE, Y. W. 2004. Crafting rules: Context-reflective data quality problem solving. *J. Manag. Inf. Syst. 20*, 3, 93–119.

LEE, Y. W., CHASE, S., FISHER, J., LEINUNG, A., MCDOWELL, D., PARADISO, M., SIMONS, J., AND YARAWICH, C. 2007a. CEIP maps: Context-Embedded information product maps. In *Proceedings of Americas' Conference on Information Systems*.

LEE, Y. W., PIERCE, E., TALBURT, J., WANG, R. Y., AND ZHU, H. 2007b. A curriculum for a master of science in information quality. *J. Inf. Syst. Educ. 18*, 2.

LEE, Y. W., PIPINO, L. L., FUND, J. F., AND WANG, R. Y. 2006. *Journey to Data Quality*. The MIT Press, Cambridge, MA.

LEE, Y. W., PIPINO, L., STRONG, D., AND WANG, R. 2004. Process embedded data integrity. *J. Database Manag. 15*, 1, 87–103.

LEE, Y. AND STRONG, D. 2004. Knowing-Why about data processes and data quality. *J. Manag. Inf. Syst. 20*, 3, 13–39.

LEE, Y., STRONG, D., KAHN, B., AND WANG, R. 2002. AIMQ: A methodology for information quality assessment. *Inf. Manag. 40*, 133–146.

LI, X. B. AND SARKAR, S. 2006. Privacy protection in data mining: A perturbation approach for categorical data. *Inf. Syst. Res. 17*, 3, 254–270.

MADNICK, S. AND PRAT, N. 2008. Measuring data believability: A provenance approach. In *Proceedings of the 41st Annual Hawaii International Conference on System Sciences*.

MADNICK, S. AND WANG, R. Y. 1992. Introduction to total data quality management (TDQM) research program. TDQM-92-01, Total Data Quality Management Program, MIT Sloan School of Management.

MADNICK, S. E., WANG, R. Y., DRAVIS, F., AND CHEN, X. 2001. Improving the quality of corporate household data: Current practices and research directions. In *Proceedings of the 6th International Conference on Information Quality*, 92–104

MADNICK, S. E., WANG, R. Y., KRISHNA, C., DRAVIS, F., FUNK, J., KATZ-HASS, R., LEE, C., LEE, Y., XIAM, X., AND BHANSALI, S. 2005. Exemplifying business opportunities for improving data quality from corporate household research. In *Information Quality*. R. Y. Wang et al., Eds. M. E. Sharpe, Armonk, NY, 181–196.

MADNICK, S. E., WANG, R. Y., AND XIAN, X. 2004. The design and implementation of a corporate householding knowledge processor to improve data quality. *J. Manag. Inf. Syst. 20*, 3, 41–69.

MADNICK, S. E. AND ZHU, H. 2006. Improving data quality with effective use of data semantics. *Data Knowl. Eng. 59*, 2, 460–475.

MARCO, D., DUATE-MELO, E., LIU, M., AND NEUHOFFAND, D. 2003. On the many-to-one transport capacity of a dense wireless sensor network and the compressibility of its data. In *Information Processing in Sensor Networks*. In Goos et al., Eds. Lecture Notes in Computer Science, vol. 2634, Springer Berlin, 556.

MIKKELSEN, G. AND AASLY, J. 2005. Consequences of impaired data quality on information retrieval in electronic patient records. *Int. J. Med. Inf. 74*, 5, 387–394.

MYERS, M. D. 1997. Qualitative research in information systems. http://www.misq.org/discovery/MISQD_isworld/index.html (retrieved on October 5, 2007).

O'CALLAGHAN, L., MISHIRA, N., MEYERSON, A., GUHA, S., AND MOTWANIHA, R. 2002. In *Proceedings of the 18th International Conference on Data and Engineering*, 685–694.

OMB (Office of Management & Budget). 2007. FEA reference models. http://www.whitehouse.gov/omb/egov/a-2-EAModelsNEW2.html. (retrieved on October 5, 2007).

ØVRETVEIT, J. 2000. The economics of quality – A practical approach. *Int. J. Health Care Quality Assurance 13*, 5, 200–207.

PETROVSKIY, M. I. 2003. Outlier detection algorithms in data mining systems. *Program. Comput. Softw. 29*, 4, 228–237.

PIERCE, E. M. 2004. Assessing data quality with control matrices. *Commun. ACM 47*, 2, 82–86.

PIPINO, L., LEE, Y., AND WANG, R. 2002. Data quality assessment. *Commun. ACM 45*, 4, 211–218.

RAGHUNATHAN, S. 1999. Impact of information quality and decision-making quality on decision quality: A theoretical model. *Decision Support Syst. 25*, 4, 275–287.

RAHM, E. AND BERNSTEIN, P. 2001. On matching schemas automatically. *VLDB J. 10*, 4, 334–350.

REDMAN, T. C. 1998. The impact of poor data quality on the typical enterprise. *Commun. ACM 41*, 2, 79–82.

SCHEKKERMAN, J. 2004. *How to Survive in the Jungle of Enterprise Architecture Frameworks: Creating or Choosing an Enterprise Architecture Framework*. Trafford Publishing.

SHANKARANARAYAN, G., ZIAD, M., AND WANG, R. Y. 2003. Managing data quality in dynamic decision environment: An information product approach. *J. Database Manag. 14*, 4, 14–32.

SHENG, Y. AND MYKYTYN, P. 2002. Information technology investment and firm performance: A perspective of data quality. In *Proceedings of the 7th International Conference on Information Quality*, 132–141.

SLONE, J. P. 2006. Information quality strategy: An empirical investigation of the relationship between information quality improvements and organizational outcomes. Ph.D. dissertation, Capella University.

STOREY, V. AND WANG, R. Y. 1998. Modeling quality requirements in conceptual database design. In *Proceedings of the International Conference on Information Quality*, 64–87

STRONG, D., LEE, Y. W., AND WANG, R. Y. 1997. Data quality in context. *Commun. ACM 40*, 5, 103–110.

TALBURT, J., MORGAN, C., TALLEY, T., AND ARCHER, K. 2005. Using commercial data integration technologies to improve the quality of anonymous entity resolution in the public sector. In *Proceedings of the 10th International Conference on Information Quality (ICIQ'05)*, 133–142.

TEJADA, S., KNOBLOCK, C., AND MINTON, S. 2001. Learning object identification rules from information extraction. *Inf. Syst. 26*, 8, 607–633.

THATCHER, M. E. AND PINGRY, D. E. 2004. An economic model of product quality and IT value. *Inf. Syst. Res. 15*, 3, 268–286.

VASSILIADIS, P., VAGENA, Z., SKIADOPOULOS, S., KARAYANNIDIS, N., AND SELLIS, T. 2001. ARKTOS: Towards the modeling, design, control and execution of ETL processes. *Inf. Syst. 26*, 537–561.

WANG, R. Y., KON, H. B., AND MADNICK, S. E. 1993. Data quality requirements analysis and modeling. In *Proceedings of the 9th International Conference of Data Engineering*, 670–677.

WANG, R. Y., LEE, Y., PIPINO, L., AND STRONG, D. 1998. Managing your information as a product. *Sloan Manag. Rev.* Summer 1998, 95–106.

WANG, R. Y. AND MADNICK, S. E. 1989. The inter-database instance identification problem in integrating autonomous systems. In *Proceedings of the 5th International Conference on Data Engineering*, 46–55.

WANG, R. Y. AND MADNICK, S. E. 1990. A polygen model for heterogeneous database systems: The source tagging perspective. In *Proceedings of the 16th VLDB Conference*, 519–538.

WANG, R. Y., REDDY, M., AND KON, H. 1995a. Toward quality data: An attribute-based approach. *Decision Support Syst. 13*, 349–372.

WANG, R. Y., STOREY, V. C., AND FIRTH, C. P. 1995b. A framework for analysis of data quality research. *IEEE Trans. Knowl. Data Eng. 7*, 4, 623–640.

WANG, R. Y. AND STRONG, D. M. 1996. Beyond accuracy: What data quality means to data consumers. *J. Manag. Inf. Syst. 12*, 4, 5–34.

WIDOM, J. 2005. Trio: A system for integrated management of data, accuracy, and lineage. In *Proceedings of the 2nd Biennial Conference on Innovative Data Systems Research (CIDR'05)*.

WINKLER, W. E. 2006. Overview of record linkage and current research directions. Tech. rep. U.S. Census Bureau, Statistics #2006-2.

XIAO, X. AND TAO, Y. 2006. Anatomy: Simple and effective privacy preservation. In *Proceedings of the 32nd VLDB Conference*.

XU H., NORD, J. H., BROWN, N., AND NORD, G. G. 2002. Data quality issues in implementing an ERP. *Industrial Manag. Data Syst. 102*, 1, 47–58.

YIN, R. 2002. *Case Study Research: Design and Methods*, 3rd ed. Sage Publications, Thousand Oaks, CA.

ZACHMAN, J. A. 1987. A framework for information systems architecture. *IBM Syst. J. 26*, 3, 276–292.

ZHU, X., KHOSHGOFTAAR, T., DAVIDSON, I., AND ZHANG, S. 2007. Editorial: Special issue on mining low-quality data. *Knowl. Inf. Syst. 11*, 2, 131–136.

Received February 2008; revised February 2009; accepted March 2009