A Framework for Analysis of Data Quality Research

Richard Y. Wang, Veda C. Storey, and Christopher P. Firth

Abstract—**Organizational databases are pervaded with data of poor quality. However, there has not been an analysis of the data quality literature that provides an overall understanding of the state-of-art research in this area. Using an analogy between product manufacturing and data manufacturing, this paper develops a framework for analyzing data quality research, and uses it as the basis for organizing the data quality literature. This framework consists of seven elements: management responsibilities, operation and assurance costs, research and development, production, distribution, personnel management, and legal function. The analysis reveals that most research efforts focus on operation and assurance costs, research and development, and production of data products. Unexplored research topics and unresolved issues are identified and directions for future research provided.**

Index Terms—**Data quality, data manufacturing, data product, Total Quality Management (TQM), ISO9000, information quality, data quality analysis, data quality practices.**

I. Introduction

For organizations to be best served by their information systems, a high degree of data quality is required, and the need to ensure this quality has been addressed by both researchers and practitioners for some time. In order to further advance the research on data quality, however, one must first understand the work that has been conducted to date and identify specific topics that merit further investigation.

The objectives of this paper are to develop a framework that can be used to analyze existing research on data quality, and to identify important future research directions. In selecting the appropriate body of research for examination, two primary criteria were used:

1) The authors of the articles specifically recognize a data quality problem, which they attempt to address in their work. That is, the research is motivated by a data quality issue.
2) The researchers address a problem that, although not specifically described as a data quality issue (e.g., user satisfaction with information systems), is comprised of components that are related to data quality management. Many of these authors' papers are referenced by research that falls into the first category.

We deliberately exclude research efforts from well established research areas that are not focused on issues directly related to data quality. For example, concurrency control has been surveyed by Bernstein and Goodman [15], data security by Denning and Denning [38], database schema integration by Batini, Lenzirini, and Navathe [14], logical database design by Teorey, Yang, and Fry [108], and temporal databases by Tansel et al. [106]. Other well established research areas that we do not investigate include privacy and software quality. Thus, this paper strives to serve as a focal point for understanding the state-of-the-art in data quality research, and to bridge the gap between research that directly addresses data quality issues and research that is primarily focused on other subject areas.

This paper is organized as follows. Section II presents a framework for data quality analysis. The elements of this framework are presented in Section III. Section IV analyzes existing research and identifies unresolved research issues for each element of the framework. Section V summarizes and concludes the paper.

II. A Framework for Data Quality Analysis

There are many approaches in the literature that can be applied to studying data quality. A data life cycle which focuses on the sequence of activities from creation to disposition of data has been proposed; see, for example, Redman [92] and Te'eni [107]. Another concept that might be applicable is the value chain, where producing, distributing and consuming organizations are categorized by the value they add at each stage in the production process [90], [91]. Other approaches to the data quality problem include an electronic data processing (EDP) audit [83], [112], [120], and database integrity [19,] [27], [28], [34], [104].

Although all of these approaches have merits, we choose to draw upon an analogy that exists between quality issues in a manufacturing environment and those in an information systems environment. Information systems have been compared to production systems wherein data are the raw materials and information (i.e., data products) is the output [29], [39]. This analogy has also been suggested in discussions of data assembly [52], information as inventory [94], information systems quality [35], and information manufacturing [4].

As shown in Fig. 1, a product manufacturing system acts on raw material inputs to produce output material, or physical products. Similarly, an information system can be viewed as a data manufacturing system acting on raw data input (e.g., a single number, a record, a file, a spreadsheet, or a report) to produce output data, or *data products* (e.g., a sorted file or a corrected mailing list). This data product, in turn, can be

Manuscript received Oct. 3, 1994.

R.Y. Wang is with the Total Data Quality Management Research Program, Room E53-320, Sloan School of Management, Massachusetts Institute of Technology, 50 Memorial Dr., Cambridge, MA 02142; e-mail: rwang@mit.edu.

V.C. Storey is with the Computer and Information Systems Dept., the William E. Simon Graduate School of Business Administration, University of Rochester, Dewy Hall 238, Rochester, NY 14627; e-mail: storey@uorgsm.bitnet.

C.P. Firth is with Global Consumer Bank, Citibank, N.A., 1 Shenton Way, Singapore 0106; e-mail: chris.firth@citicorp.com.

IEEECS Log Number K95053.

1041-4347/95$04.00 © 1995 IEEE

treated as raw data in another data manufacturing system. Use of the term "data manufacturing" encourages researchers and practitioners alike to seek out cross-disciplinary analogies that can facilitate the transfer of knowledge from the field of product quality to the field of data quality. Use of the term "data product" emphasizes the fact that the data output has value that is transferred to customers, be they internal or external to the organization.

	Product Manufacturing	Data Manufacturing
Input	Raw Materials	Raw Data
Process	Materials Processing	Data Processing
Output	Physical Products	Data products

Fig. 1. An analogy between physical products and data products.

The preceding considerations led us to select the International Organization for Standardization's ISO9000 [54] for our analysis. The objectives of the ISO9000 are:

1) to clarify the distinctions and interrelationships among the principal quality concepts, and
2) to provide guidelines for the selection and use of a series of International Standards on quality systems that can be used for internal quality management purposes (ISO9004) and for external quality assurance purposes (ISO9001, ISO9002, and ISO9003).

For convenience, the term ISO9000 is used hereafter to refer to the 9000 series (ISO9000 to ISO9004 inclusive). The main strength of the ISO approach is that it is a set of well-established standards and guidelines that has been widely adopted by the international community. It provides guidance to all organizations for quality management purposes, with a focus on the technical, administrative and human factors affecting the quality of products or services, at all stages of the quality loop from detection of need to customer satisfaction. An emphasis is placed on the satisfaction of the customer's needs, the establishment of functional responsibilities, and the importance of assessing (as far as possible) the potential risks and benefits. All of these aspects are considered in establishing and maintaining an effective quality system.

Terms and definitions used in the ISO9000 are described in the ISO8402. By rephrasing the five key terms given for product quality in the ISO8402, we define the following terms that are needed to develop this present framework:

- A *data quality policy* refers to the overall intention and direction of an organization with respect to issues related to the quality of data products. This policy is formally expressed by top management.
- *Data quality management* is the management function that determines and implements the data quality policy.
- A *data quality system* encompasses the organizational structure, responsibilities, procedures, processes, and resources for implementing data quality management.
- *Data quality control* is the set of operational techniques and activities that are used to attain the quality required for a data product.
- *Data quality assurance* includes all those planned and systematic actions necessary to provide adequate confidence that a data product will satisfy a given set of quality requirements.

A framework for data quality analysis is developed that consists of seven elements (as shown in Fig. 2) adapted from the ISO9000: 1) management responsibilities, 2) operation and assurance costs, 3) research and development, 4) production, 5) distribution, 6) personnel management, and 7) legal function. These seven are a result of grouping the original twenty categories in the ISO9004 to obtain a workable number of dimensions that could be applied to data quality.

III. Elements of the Framework

This section describes each of the seven elements in this framework.

A. Management Responsibilities

Top management is responsible for developing a corporate data quality policy. This policy should be implemented and maintained to be consistent with other policies, especially other quality policies. Management should also identify the company's critical data quality requirements and establish a data quality system that applies to all phases of the production of data products.

B. Operation and Assurance Costs

Unlike most raw materials, data are not consumed when processed and, therefore, may be reused repeatedly. Although the cost of data "waste" may be negligible, the cost of using inaccurate data certainly may be large [13], [88]. The impact of data product quality can be highly significant, particularly in the long term. It is, therefore, important that the costs of a data quality system be regularly reported to, and monitored by, management, and related to other cost management. These costs can be broadly divided into operating costs and assurance costs. Operating costs include prevention, appraisal, and failure costs. Assurance costs relate to the demonstration and proof required by customers and management.

C. Research and Development

Those responsible for research and development should work with those responsible for marketing to establish quality requirements for data products, whether they will be distributed to internal or to external customers of the organization. Together, they should determine the need for a data product, define the market demand (either internal or external), and determine customer requirements regarding the quality of the data product.

The specification and design function translates the quality requirements of data products into technical specifications for

	Element	Description
1.	Management Responsibilities	• Development of a corporate data quality policy • Establishment of a data quality system
2.	Operation and Assurance Costs	• Operating costs include prevention, appraisal, and failure costs • Assurance costs relate to the demonstration and proof of quality as required by customers and management
3.	Research and Development	• Definition of the dimensions of data quality and measurement of their values • Analysis and design of the quality aspects of data products • Design of data manufacturing systems that incorporate data quality aspects
4.	Production	• Quality requirements in the procurement of raw data, components, and assemblies needed for the production of data products • Quality verification of raw data, work-in-progress, and final data products • Identification of non-conforming data items and specifications of corrective actions
5.	Distribution	• Storage, identification, packaging, installation, delivery, and after-sales servicing of data products • Quality documentation and records for data products
6.	Personnel Management	• Employee awareness of issues related to data quality • Motivation of employees to produce high-quality data products • Measurement of employee's data quality achievement
7.	Legal Function	• Data product safety and liability

Fig. 2. A framework for data quality research.

the raw data, the data manufacturing process, and the data products themselves. This function should be such that the data product can be produced, verified, and controlled under the proposed processing conditions. The quality aspects of the design should be unambiguous and adequately define characteristics important to data product quality, such as acceptance and rejection criteria. Tests and reviews should be conducted to verify the design.

D. Production

In terms of production, data quality begins with the procurement of raw data. A data quality system for procurement must include the selection of qualified raw data suppliers, an agreement on quality assurance, and the establishment of verification methods. All raw data must conform to data quality standards before being introduced into the data manufacturing system. When traceability is important, appropriate identification should be maintained throughout the data manufacturing process to ensure that the original identification of the raw data and its quality status can be attained. The identification, in the form of tags, should distinguish between verified and unverified raw data. Verification of the quality status of data should be carried out at important points during the data manufacturing process. The verifications should relate directly to the data product specifications.

Suspected nonconforming data items (including raw data, data products in-process, and final data products) should be identified, reviewed, and recorded. Corrective action begins when a nonconforming data item is detected. The significance of a problem affecting quality should be evaluated in terms of its potential impact on production costs, quality costs, customer satisfaction, and so forth. Cause and effect relationships should be identified and preventative action initiated to prevent a reoccurrence of nonconformity. It is well accepted by practitioners in the data quality management area that it is more beneficial to check and, if appropriate, modify the process that caused the data quality problem than to correct the nonconforming data items.

E. Distribution

The handling of data products requires proper planning and control. The marking and labeling of data products should be legible and remain intact from initial receipt to final delivery. Specifically, in the context of data stored in computerized databases, this means that the metadata that describes the data should be interpretable to the user and not corrupted. The data quality system should establish, and require the maintenance and after-sales servicing of data products, whether they be distributed to internal or external customers. In addition, it should provide a means of identifying, collecting, indexing, storing, retrieving, and disposing of documentation and records regarding the data products produced by a data manufacturing system.

F. Personnel Management

Personnel management falls into three categories: training, qualification (formal qualification, experience, and skill), and motivation. An effort should be made to raise employees' awareness of issues related to data quality. Management should measure and reward data quality achievement.

G. Legal Function

The safety aspect of data products should be identified in order to enhance product safety and minimize product liability. Doing so includes identifying relevant safety standards, carrying out design evaluation and prototype testing, providing data quality indicators to the user, and developing a traceability system to facilitate data product recall.

IV. Analysis of Data Quality Research

The framework is employed in this section to analyze articles relevant to data quality research. In addition to the articles with which the authors were already familiar, exhaustive online database searches were conducted. Since we were aware that research on data quality may appear in different disciplines, we searched various databases, including Inspec and ABI/Inform; the former focuses on computing literature and the later contains business, management, and industry-specific articles.

The Inspec library that we searched contains articles dating back to 1970. Our search included all the articles in the journal and conference proceedings published by the Association of Computing Machinery (ACM) and by the Institute of Electrical and Electronics Engineers (IEEE). Any phrase containing the words "data" (or "information") and "quality" adjacent to each other with no more than two words in between was used as a key phrase in this search. For example, any article with a title, key phrase, or abstract that contains the words "data quality," "quality of information," or "data and process quality" was identified. This exhaustive search found relatively few articles with the key phrases specified above. It did, however, identify all the research papers of which we were previously aware. Another on-line search was conducted using other key words such as "integrity," "accuracy," and "consistency."

The ABI/Inform library that we searched contains articles published between January 1987 and April 1993. Since this database system does not have the capability to search for two words adjacent to each other with no more than two words in between, we used a similar strategy where articles with the words "data" (or "information") and "quality" in the title, abstract, or keywords were identified. Major journals related to the information systems field were searched, including *Decision Sciences, Decision Support Systems, Harvard Business Review, Management Science, MIS Quarterly*, and *Sloan Management Review*. (The ABI/Inform Library that we researched does not include *Information Systems Research* (ISR), the *Journal of Management Information Systems* (JMIS), and conference proceedings in the information systems field. Therefore, a manual search of these journals was conducted.) Other on-line databases that were used to identify articles include Compendex which contains a broad spectrum of engineering articles, and Lexis/Nexis which contains a variety of articles including legal matters and news stories.

Articles that met the criteria described at the beginning of this paper were collected and their references further examined to uncover additional relevant papers. Although many articles identified above could have useful implications for data quality, they are not included in this paper. For example, much research in the areas of concurrency control, integrity constraints, and temporal databases is aimed at ensuring the consistency, accuracy, and currency of the stored data, but is not included.

This section is divided into seven subsections reflecting the seven elements of the framework:

A. Management Responsibilities
B. Operation and Assurance Costs
  B.1 Information Systems
  B.2 Database
  B.3 Accounting
C. Research and Development
  C.1 Analysis and Design of the Quality Aspects of Data Products
  C.2 Incorporating Data Quality into the Design of Data Manufacturing Systems
  C.3 Dimensions of Data Quality and Measurement of Their Values
D. Production
E. Distribution
F. Personnel Management
G. Legal Function

To serve as a reference for the reader, related research efforts are listed in a table at the beginning of each subsection. The tables also reflect whether these research efforts address issues related to other elements of the framework. For example, Table I in Section IV.A shows that Huh et al. [52] also address issues related to Sections IV.C.3 and IV.D, as indicated by the symbol √. This convention is used throughout Section IV.

A. Management Responsibilities

The importance of top management's involvement in attaining high quality data has been recognized; see, for example, Bailey [6] and Halloran et al. [48]. However, as can be seen from Table I, very little research has been conducted to investigate what constitutes a data quality policy or how to establish a data quality system, both of which are important tasks for top management.

Ballou and Pazer [10] present quantitative measurements for analyzing errors in conjunctive, multi-criteria, satisficing decision processes. Their results support ideas previously couched in qualitative terms only. Oman and Ayers [86] focus on the organizational side of implementing data quality systems. They use data quality survey results as a tool to improve the quality of data in the U.S. government. The survey results were sent back to the people directly responsible for data quality on a monthly basis over a period of several years.

AT&T has taken a pragmatic approach to data quality management that is based on its process management and improvement guidelines. The approach involves seven steps: 1) establishing a process owner, 2) describing the current process and identifying customer requirements, 3) establishing a system of measurement, 4) establishing statistical control, 5) identifying improvement opportunities, 6) selecting improvement opportunities to pursue, and setting objectives for doing

TABLE I
DATA QUALITY LITERATURE RELATED TO MANAGEMENT RESPONSIBILITIES

Section	IV.A		IV.B			IV.C		IV.D	IV.E	IV.F	IV.G
Research		IV.B.1	IV.B.2	IV.B.3	IV.C.1	IV.C.2	IV.C.3				
[Ballou & Pazer, 1990]	✓										
[Huh, Keller, Redman & Watkins, 1990]	✓						✓	✓			
[McGee, 1992]	✓										
[Oman & Ayers, 1988]	✓							✓		✓	
[Pautke & Redman, 1990]	✓						✓	✓			
[Redman, 1992]	✓						✓	✓			
[Redman, 1995]	✓										
[Wang & Kon, 1993]	✓										

so, and 7) making changes and sustaining gains. This approach has been successfully applied both within and outside AT&T [52], [89], [92], [93].

Another approach to data quality management entails two phases and is found in many organizations [78], [115]. In the first phase, the data quality proponent initiates a data quality project by identifying an area in which the effectiveness of an organization is critically impacted by poor data quality. Upon completion of this phase, the proponent has gained experience and established credibility within the organization. In the second phase, the proponent strives to become the leader for data quality management within the organization.

A high level management perspective of data quality has also been proposed [78], [114] in which fundamental problems such as the need to define, measure, analyze, and improve data quality are identified. Guidelines are provided to show how total data quality management may be attained. In addition, challenging research issues for each problem area are identified.

Other approaches have also been developed by various firms. Although the implementation details vary with different approaches and organizational settings, most practitioners have benefited from the accumulated body of knowledge on total quality management (TQM) [32], [37], [40], [63], [105].

A.1. Analysis of Research.

The lack of research on what constitutes a data quality policy and a data quality system contrasts sharply with the growing anecdotal evidence that organizations are increasingly aware of the need to develop a corporate policy for data quality management [20], [30], [45], [71], [72], [97]. Researchers and practitioners alike need to demonstrate convincingly to top management that data quality is critical to the survival of an organization in the rapidly changing global environment. Research is also needed to develop methodologies that will assist management in identifying data quality factors that affect a company's position. Such methodologies will allow management to analyze the impact of data quality more directly. A useful starting point might be to update the questionnaire developed by Saraph, Benson, and Schroeder [95] that measures management policies related to total quality management.

Case studies that document successful corporate data quality policies and data quality systems are needed. From them, hypotheses for a corporate data quality policy and for system guidelines could be developed. As more and more companies implement corporate data quality policies and systems, empirical analyses could identify the most successful approaches. The results from the empirical analyses would help to develop a body of knowledge on this topic.

B. Operation and Assurance Costs

The costs of a data quality system are divided into: 1) operating costs (prevention, appraisal, and failure), and 2) assurance costs (demonstrating the quality required by both management and customers). The research that addresses cost issues is identified in Table II. Although the literature does not distinguish between operating and assurance costs, it is possible to classify the work that has been carried out by the areas in which the work occurs, namely, information systems (Section IV.B.1), databases (Section IV.B.2), and accounting (Section IV.B.3).

B.1. Information Systems

The information systems literature provides cost and quality tradeoffs for control procedures in information systems and a methodology for allocating resources to data quality enhancement. Ballou and Pazer [9] present a framework for studying the cost/quality tradeoffs of internal control scenarios designed to ensure the quality of the output from an information system. The model includes the following parameters: the cost and quality of processing activities, the cost and quality of corrective procedures, and a penalty cost that is incurred by failure to detect and correct errors. Ballou and Tayi [12] use parameters that include the cost of undetected errors, the stored data error rate, and the effectiveness of data repair procedures in an integer program designed to allocate available resources in an optimal manner. The researchers note in particular, the need to obtain better estimates for the penalty costs of poor data quality. A heuristic is developed that helps assess whether it is better to identify and correct few, serious errors in one database or more widespread, but less severe, quality problems in another database.

TABLE II
DATA QUALITY LITERATURE RELATED TO OPERATION AND ASSURANCE COSTS

Section	IV.A		IV.B			IV.C		IV.D	IV.E	IV.F	IV.G
Research		IV.B.1	IV.B.2	IV.B.3	IV.C.1	IV.C.2	IV.C.3				
[Amer, Bailey & De, 1987]				✓							
[Ballou & Pazer, 1987]		✓				✓	✓				
[Ballou & Tayi, 1989]		✓					✓	✓			
[Bodnar, 1975]				✓							
[Bowen, 1993]				✓			✓			✓	
[Burns & Loebbecke, 1975]				✓							
[Cushing, 1974]				✓							
[Feltham, 1968]				✓			✓				
[Fields, Sami & Sumners, 1986]				✓							
[Groomer & Murthy, 1989]				✓							
[Hamlen, 1980]				✓							
[Hansen, 1983]				✓							
[Johnson, Leitch & Neter, 1981]				✓							
[Mendelson & Saharia, 1986]			✓								
[Nichols, 1987]				✓							
[Stratton, 1981]				✓							
[Wand & Weber, 1989]				✓							
[Yu & Neter, 1973]				✓							

B.2. Databases

Database research has been conducted that investigates the cost and impact of incomplete information on database design. Mendelson and Saharia [81] present a decision-theoretic approach to doing so. An implicit assumption in this work is that the database will reflect the real world as far as the relevant schemas are concerned; in other words, for each schema, the database will contain the complete set of instances at any given point in time and all the instance values will be accurate. The premise of this research is that a database designer must make a tradeoff between the cost of incomplete information (i.e., the exclusion of relevant entities or attributes) and data-related costs (i.e., the cost of data collection, manipulation, storage, and retrieval).

A cumulated body of research has also been developed in the database area that would reduce CPU usage time and query response time. Other research aims at minimizing communication costs, with a goal of increasing data availability [67]. Although this type of research can clearly be associated with operation costs, for reasons previously mentioned, it is not discussed further.

B.3. Accounting

A body of research in the accounting literature places specific emphasis on internal control systems and audits. An implicit assumption underlying this research is that by incorporating a set of internal controls into a financial information system to enhance the system's reliability, it will be possible to maintain a high probability of preventing, detecting, and eliminating data errors, irregularities, and fraud. The demonstrated reliability of the system can provide evidence of the quality of the data products produced by the system.

Feltham [42] identifies relevance, timeliness, and accuracy as the three dimensions of data quality, and analyzes their relationship with the value-in-excess-of-cost criterion, within the context of the data consumer. A payoff function is developed in terms of future events and the prediction of these events. Yu and Neter [122] propose one of the earliest stochastic models of an internal control system that could serve as the basis for an objective, quantitative evaluation of the reliability of an internal control system. They also discuss implementation problems for the proposed model and possible approaches to tying the output from the proposed model to substantive tests of an account balance. Cushing [33] independently develops a simple stochastic model for the analysis and design of an internal control system by adapting concepts from the field of reliability engineering. The model uses reliability and cost parameters to provide a means of computing the reliability of a process, that is, the probability of completion with no errors. The cost parameters include the cost of performing the control procedure, the average cost of searching and correcting for signaled errors, and the average cost of an undetected error.

Burns and Loebbecke [23] focus on internal control evaluation in terms of a tolerable data error compliance level, but do not address issues related to cost. Bodnar [17] expands Cushing's discussion of the applicability of reliability engineering techniques to internal control systems analysis. Hamlen [50] proposes a model for the design of an internal control system that minimizes system costs subject to management-set error reduction probability goals for certain types of errors. Stratton [99] demonstrates how reliability models can be used by management, or independent auditors, to analyze accounting internal control systems.

Other research investigates error characteristics [18], [60], audit concerns in distributed processing systems [51], models of the internal control process [43], [85], information systems research as applied to accounting and auditing [3], data quality as it relates to audit populations [47], and how auditors need to examine only those parts of a system where structural changes made at the system level induce structural changes at the subsystem level [112].

B.4. Analysis of Research.

A significant body of research exists that addresses issues associated with costs and data quality, mostly data errors. At a theoretical level, this body of research provides a basis for studying the relationship between prevention costs and failure costs. At a more pragmatic level, however, there is still a critical need for the development of methodologies and tools that help practitioners to determine operation and assurance costs when implementing a data quality system. Such methodologies and tools should allow practitioners to determine prevention, appraisal, and failure costs along data quality dimensions such as accuracy and timeliness.

There does not appear to be any documented research dealing with the economics of external failure, although anecdotes abound; see, for example, Liepins and Uppuluri [72]. Furthermore, there does not seem to be any work that attempts to estimate the costs of data quality assurance. Applying quality control techniques to data has been largely overlooked because the economic consequences of data errors are less apparent than are manufacturing non-conformities. But data errors are costly–the cost is simply harder to quantify [71]. In some cases, it can be catastrophically high. (In a 1990 report to Sen. Sam Nunn (D-Ga.), the U.S. General Accounting Office reported that more than $2 billion of federal loan money had been lost because of poor data quality at a single agency.)

In addition, persuasive case studies are needed of the impact of data quality upon profit and loss statements, similar to those that have been conducted in the physical product manufacturing area. For example, when examining product or service quality, Crosby [31] estimates that manufacturing companies spend over 25% of sales doing things wrong and service companies spend more than 40% of their operating costs on wasteful practices. In another example, Garvin [46] demonstrates that among businesses with less than 12% market share, those with inferior product quality averaged a return on investment (ROI) of 4.5%, those with average product quality an ROI of 10.4%, and those with superior product quality, an ROI of 17.4%. Those businesses that improved their product quality during the 1970s increased their market share five to six times faster than those whose product quality declined—and three times faster than those whose quality remained unchanged. Research in the data quality area analogous to the examples illustrated above will contribute not only to the development of tools and methodologies in estimating operation and assurance costs, but will also help convince top management to implement a corporate data quality policy.

C. Research and Development

A significant amount of work can be classified under research and development, although the original work might not have been identified as such. From the data quality management perspective, there are three main issues involved in research and development: analysis and design of the data quality aspects of data products (Section IV.C.1), design of data manufacturing systems that incorporate data quality aspects (Section IV.C.2), and definition of data quality dimensions and the measurement of their values (Section IV.C.3). The following three subsections examine the literature, as summarized in Table III, that addresses these issues.

C.1. Analysis and Design of the Quality Aspects of Data Products

Brodie [19] places the role of data quality within the lifecycle framework with an emphasis on database constraints. Data quality is defined as the extent to which a database accurately represents the essential properties of the intended application, and has three distinct properties: 1) data reliability, 2) logical or semantic integrity, and 3) physical integrity (the correctness of implementation details). A semantic integrity subsystem to improve data quality is proposed that consists of five parts: 1) a constraint language to express the constraints, 2) a constraint verifier, 3) a constraint database management system, 4) a constraint validation system, and 5) a violation-action processor.

Svanks [104] reports on the actual development of an integrity analysis system that has been tested on a case study. Svanks' approach consists of seven steps: 1) defining database constraints, 2) selecting statistical techniques for sampling the database, 3) selecting the integrity analysis to be performed, 4) defining suitable quality measures, 5) specifying outputs to be produced from a defect file, 6) developing and testing program code, and 7) executing the integrity analysis system.

In the conceptual modeling area, most data models, including the entity-relationship (ER) model [25], are aimed at capturing the content of data (such as which entities or attributes are to be included for the intended application domain) and do not deal explicitly with the data quality aspect. Chen, who first proposed the ER model, recommends that a methodology be developed to incorporate quality aspects into the ER model [26]. To extend the ER model, a methodology for data quality requirements collection and documentation is proposed to include data quality specifications as an integral component of the database design process [115]. The methodology includes a step-by-step procedure for defining and documenting the data quality parameters (e.g., timeliness or credibility) that are important to users. The subjective data quality parameters are then translated into more objective data quality indicators (e.g., data source, creation time, and collection method) that should be tagged to the data items.

C.1.a. Analysis of Research. Previous research has focused primarily on the accuracy requirements for data products that are represented by semantic integrity constraints. Semantic integrity constraints, however, do not capture all the accuracy requirements. For example, the zip code 02146 corresponds to

TABLE III
DATA QUALITY LITERATURE RELATED TO RESEARCH AND DEVELOPMENT

Section	IV.A		IV.B			IV.C		IV.D	IV.E	IV.F	IV.G
Research		IV.B.1	IV.B.2	IV.B.3	IV.C.1	IV.C.2	IV.C.3				
[Agmon & Ahituv, 1987]							✓				
[Ahituv, 1980]							✓				
[Bailey & Pearson, 1983]	✓						✓				
[Ballou & Pazer, 1982]							✓				
[Ballou & Pazer. 1985]					✓	✓	✓				
[Ballou & Pazer, 1987]		✓				✓	✓				
[Ballou & Tayi, 1989]		✓					✓	✓			
[Ballou, Wang, Pazer & Tayi, 1993]						✓	✓				
[Ballou & Pazer, 1995]						✓					
[Blaylock & Rees, 1984]							✓				
[Bowen, 1993]				✓			✓				
[Brodie, 1980]					✓			✓			
[Chen, 1993]					✓						
[Delone & McLean, 1992]							✓				
[Feltham, 1968]				✓			✓				
[Halloran et al., 1978]	✓						✓				
[Hamilton & Chervany, 1981]							✓				
[Huh, Keller, Redman & Watkins, 1990]	✓						✓	✓			
[Iivari & Koskela, 1987]							✓				
[Ives, Olson & Baroudi, 1983]							✓				
[Ives & Olsen, 1984]							✓				
[Jang, Ishii & Wang, 1995]						✓	✓	✓			
[Janson, 1988]							✓	✓			
[Jones & McLeod, 1986]							✓				
[Kim, 1989]							✓				
[King & Epstein, 1983]							✓				
[Kriebel, 1979]							✓				
[Larcker & Lessig, 1980]							✓				
[Melone, 1990]							✓				
[Miler & Doyle, 1987]							✓				
[Page & Kaomea, 1994]						✓		✓	✓		
[Morey, 1982]							✓	✓			
[Paradice & Fuerst, 1991]							✓	✓			
[Pautke & Redman, 1990]	✓						✓	✓			
[Redman, 1992]	✓						✓	✓			
[Strong, Lee & Wang, 1994]							✓				
[Svanks, 1984]					✓			✓			
[Wand & Wang, 1994]							✓				
[Wang & Madnick, 1990]						✓		✓			
[Wang, Kon & Madnick, 1993]					✓	✓		✓			
[Wang, Reddy & Kon, 1992]						✓					
[Wang, Reddy & Gupta, 1993						✓	✓	✓	✓		
[Wang, Strong & Guarascio, 1994]							✓				
[Zmud, 1978]							✓				

Brookline, Massachusetts, except for Chestnut Hill (02167). A semantic integrity constraint provided by the database system may not be able to capture this exception. Since data quality is a multi-faceted concept that includes not only accuracy, but also other dimensions such as timeliness and completeness, much more research is needed on the other dimensions as well. In short, research on this topic is still in its formative stage.

Much work is needed to develop a formal design methodology that can be used by database designers to systematically gather and translate "soft" customer data quality requirements into "hard" design specifications. Research issues such as the following need to be addressed:

1) What differentiates a data quality attribute from a regular entity attribute?
2) How does one relate quality attributes to entities, relationships, and their attributes?
3) How does one determine which quality attributes are appropriate for a given application domain?
4) Under what circumstances should an attribute be categorized as a quality attribute as opposed to an entity attribute?
5) What are the criteria for trading off the quality aspects versus other design factors, such as cost, when determining which quality attributes to incorporate into a database system?

C.2. Incorporating Data Quality into the Design of Data Manufacturing Systems

Similar to the quality-by-design concept that is advocated by leaders in the TQM area [63], [105], the quality aspects of data products should be designed into data manufacturing systems in order to attain quality-data-product-by-design. The literature that addresses this topic can be classified into two categories: 1) analytical models that study how data manufacturing systems can be developed to meet data quality requirements (e.g., acceptable error rate) subject to certain constraints (e.g., minimal cost), and 2) designing system technologies into data manufacturing systems to ensure that data products will meet the specified quality.

C.2.a. Analytical Models. Most of the research that addresses issues related to operation and assurance costs falls into this category. This is because the research approaches are analytic, and their primary concern is how to enhance a data manufacturing system's quality in such a way that a high probability of preventing, detecting, and eliminating data quality problems in the system can be maintained.

In addition, Ballou and Pazer [8], [9] describe a model that produces expressions for the magnitude of errors for selected terminal outputs. The model is intended for analysts to use in comparing alternative quality control strategies and is useful in assessing the impact of errors in existing systems. The researchers also develop an operations research model for analyzing the effect and efficacy of using data quality control blocks in managing production systems. Ballou et al. [13] further propose a data manufacturing model to determine data product quality. A set of ideas, concepts, models, and procedures appropriate to data manufacturing systems is presented that can be used to assess the quality impact of data products delivered or transferred to data customers. To measure the timeliness, quality, and cost of data products, the model systematically tracks relevant parameters. This is facilitated through a *data manufacturing analysis matrix* that relates data units to various system components.

C.2.b. Systems Technologies. A *data tracking* technique that employs a combination of statistical control and manual identification of errors and their sources has been developed [52], [89], [92]. Underlying the data tracking technique is the idea that processes that create data are often highly redundant. Data tracking uses the redundancy to determine pairs of steps in the overall process that yield inconsistent data. Changes that arise during data tracking are classified as normalization, translation, or spurious-operational. Spurious-operational changes occur when fields are changed during one sub-process; they indicate an error somewhere in the process. This allows the cause of errors to be systematically located.

An attribute-based model that can be used to incorporate quality aspects of data products has also been developed [57], [115], [116], [117], [119]. Underlying this model is the idea that objective data quality indicators (such as source, time, and collection method) can be designed into a system in such a way that they will be delivered along with the data products. As a result, data consumers can judge the quality of the data product according to their own chosen criteria. These data quality indicators also help trace the supplier of raw data to ensure that the supplier meets the quality requirements. Also introduced in the attribute-based model is the notion of a *quality database management system*; that is, one that supports data quality related capabilities (for example, an SQL that includes a facility for data quality indicators). Much like the entity-relationship model [25] which has been widely adopted by the industry as a tool for database design, the attribute-based model can be used as a guideline by database designers in incorporating data quality aspects into their systems.

Finally, research by Brodie [19] and Svanks [104] on database constraints could be further extended to help design and produce data products that will meet the specified quality.

C.2.c. Analysis of Research. A significant number of analytic models have been developed for the design of data manufacturing systems, most of which focus primarily on data accuracy. Future research in this area should be directed toward other data quality dimensions. Ballou and Pazer [11] define the *accuracy-timeliness tradeoff* as the balance between using current but inaccurate information or accurate but outdated information. Based on the definition, they analyzed a generic family of environments, and procedures are suggested for reducing the negative consequences of this tradeoff. Their research indicates that in many situations, rather general knowledge concerning relative weights and shapes of functions is sufficient to determine optimizing strategies.

Many of these mathematical models make assumptions that require further work in order to be applicable in practice. The assumptions underlying these models typically include inputs for the mathematical models, the topology of the system, the cost and data quality information, and the utility function of the data consumers. In practice, obtaining this information can be very challenging, and there is also a gap between understanding these models and applying them to the design of industrial-strength data manufacturing systems.

To fully exploit the potential of these mathematical models, computer-aided tools and methodologies based on extensions to them need to be developed. They would allow designers to more systematically explore the design alternatives for data manufacturing systems, much like the computer-aided software engineering tools that have been developed, based on variants of the entity-relationship model, for database designers to explore design alternatives.

System technologies research represents one of the promising areas that can have short-term as well as long-term benefits to organizations. For example, Page and Kaomea [87] present a system which will be deployed on U.S. aircraft carriers as a stand-alone image exploitation tool, particularly for making trade-offs between the timeliness of receiving a tactical image vs. the degree of accuracy of the image. In the long run, techniques for making trade-offs among data quality dimensions should be integrated with other efforts, with the goal of exploiting the data tracking technique, the attribute-based approach, and the database constraints work.

C.3. Dimensions of Data Quality and Measurement of Their Values

The three primary types of researchers who have attempted to identify appropriate dimensions of data quality are those in the areas of: 1) data quality, 2) information systems success and user satisfaction, and 3) accounting and auditing.

In the data quality area, a method is proposed by Morey [84] that estimates the "true" stored error rate. Ballou et al. [7], [8], [9], [12], [13] define: 1) accuracy which occurs when the recorded value is in conformity with the actual value, 2) timeliness which occurs when the recorded value is not out of date, 3) completeness which occurs when all values for a certain variable are recorded, and 4) consistency which occurs when the representation of the data value is the same in all cases. Strong, Lee, and Wang [102], and Wang, Strong, and Guarascio [118] identify intrinsic, contextual, representation, and accessibility aspects of data as four categories for data quality. Other dimensions that have been identified include data validation, availability, traceability, and credibility [58], [74], [116]. Redman [92] identifies more than 20 dimensions of data quality, including accuracy, completeness, consistency, and cycle time. Finally, Paradice and Fuerst [88] develop a quantitative measure of data quality by formulating the error rate of MIS records, which are classified as being either "correct" or "erroneous."

A cumulated body of research has appeared in the information systems field based on evaluating information systems success from the user's point of view. Halloran et al. [48] propose various factors such as usability, reliability, independence, and so forth. Zmud [123] conducts a survey to establish important user requirements of data quality. The results of Zmud's work reveal some of the users' intuition about the dimensions of data quality. In evaluating the quality of information systems, Kriebel [66] identifies attributes such as accuracy, timeliness, precision, reliability, completeness, and relevancy. In assessing the value of an information system, Ahituv [2] proposes a multi-attribute utility function, and suggests relevant attributes such as timeliness, accuracy, and reliability.

User satisfaction studies have identified as important dimensions accuracy, timeliness, precision, reliability, and completeness [5]. Other work on user satisfaction and user involvement that identifies data quality attributes can be found in Ives, Olson, and Baroudi [56], Ives and Olson [55], Kim [64], and Melone [80]. Work also has been carried out on information systems value [36], [65].

Agmon and Ahituv [1] apply reliability concepts from the field of quality control to information systems. Three measures for data reliability are developed: 1) internal reliability (the "commonly accepted" characteristics of data items), 2) relative reliability (the compliance of data to user requirements), and 3) absolute reliability (the level of resemblance of data items to reality). Jang, Ishii, and Wang [57] propose a data quality reasoner that provides both an automated form of judging data quality and an objective measure of overall data quality.

Other, perhaps less directly related research includes: development of an instrument for perceived usefulness of information [68], analysis of approaches for evaluating system effectiveness [49], examination of the "usefulness" of information in relationship to cognitive style [16], evaluation of the structure of executive information systems and their relationship to decision making [62], measurement of the quality of information systems [53], and measurement of the effectiveness of information systems as applied to financial services [82].

In accounting and auditing, where internal control systems require maximum reliability with minimum cost, the key data quality dimension used is *accuracy*–defined in terms of the frequency, size, and distribution of errors in data. In assessing the value of information, Feltham [42] further identifies *relevance* and *timeliness* as desirable attributes of information.

C.3.a. Analysis of Research. A number of data quality dimensions have been identified, although there is a lack of consensus, both on what constitutes a set of "good" data quality dimensions, and on what an appropriate definition is for each. In fact, even a relatively obvious dimension such as accuracy, does not have a well established definition. Most of the research efforts have assumed that a record is accurate if it conforms with the actual value. This, however, would lead one to ask several questions:

- What is the "actual value?"
- Do values in all fields of a record need to conform with the "actual value" in order to be considered as "accurate?"
- Should a record with one inaccurate field value be defined as more accurate than a record that has two inaccurate field values?
- Would a file with all of its records 99% accurate (1% off from the actual value) be more accurate than a file with 99% of its records 100% accurate but 1% of its records off by an order of magnitude?

Two avenues could be pursued in the establishment of data quality dimensions. The first is to use a scientifically grounded approach to rigorously define dimensions of data quality, and to separate the dimensions into those intrinsic to an information system from those external to the system. An ontological-based approach [21], [22], [110], [111], [113], for example, identifies the data deficiencies that exists when mapping the real world to an information system, and therefore, offers a rigorous basis for the definition of dimensions of data quality. Another approach is to apply information theory as a basis for these dimensions [36]. Marketing research is yet another ap-

TABLE IV
DATA QUALITY LITERATURE RELATED TO PRODUCTION

Section	IV.A		IV.B			IV.C		IV.D	IV.E	IV.F	IV.G
Research		IV.B.1	IV.B.2	IV.B.3	IV.C.1	IV.C.2	IV.C.3				
[Ballou & Tayi, 1989]		✓					✓	✓			
[Brodie, 1980]					✓			✓			
[Fellegi & Holt, 1976]								✓			
[Garfinkel, Kunnathur & Liepens, 1986]								✓			
[Huh, Keller, Redman & Watkins, 1990]	✓						✓	✓			
[Jang, Ishii & Wang, 1995]						✓	✓	✓			
[Janson, 1988]							✓	✓			
[Jaro, 1985]								✓			
[Liepens, Garfinkel & Kunnathur, 1982]								✓			
[Little & Smith, 1987]								✓			
[McKeown, 1984]								✓			
[Morey, 1982]							✓	✓			
[Oman & Ayers, 1988]	✓							✓		✓	
[Page & Kaomea, 1994]						✓		✓	✓		
[Paradice & Fuerst, 1991]							✓	✓			
[Pautke & Redman, 1990]	✓						✓	✓			
[Redman, 1992]	✓						✓	✓			
[Svanks, 1984]					✓			✓			
[Strong, 1988]								✓			
[Strong, 1993]								✓			
[Strong & Miller, 1993]								✓			
[Wang & Madnick, 1990]						✓		✓			
[Wang, Kon & Madnick, 1993]					✓			✓			
[Wang, Reddy & Gupta, 1993]						✓	✓	✓	✓		

proach to empirically derive and define dimensions of importance to data consumers [118]. This approach would develop a framework that captures the aspects of data quality that are important to data consumers.

The second avenue is to establish pragmatic approaches for defining data quality in an operational manner. One approach would be to have data quality defined by the user, depending on the context of the task at hand. Alternatively, it may be useful to form a data quality standard technical committee, consisting of key participants from government, industry, and research institutions. The responsibility of this committee would be to recommend a set of operational definitions for data quality dimensions. For example, the IEEE has developed a standard for software quality dimensions [76], [96].

D. Production

The previous section dealt with issues related to the research and development of data manufacturing systems that will enable the data producer to manufacture data products with the specified quality demanded by the data consumer. This section focuses on how to ensure that a data product is manufactured according to its given data quality specifications. In producing data products, three main issues are involved: 1) quality requirements in the procurement of raw data, components, and assemblies needed for the production of data products, 2) quality verification of raw data, work-in-progress, and final data products, and 3) nonconformity, and corrective action for data products that do not conform to their specifications.

Table IV summarizes the data quality research related to production. Most of the work on the research and development of data products can be employed at production time to help understand how to address the inter-related issues of procurement, verification, nonconformity, and corrective action. In addition, many research efforts address these issues directly at production time, as discussed below.

Morey [84] focuses on applications that are updated periodically, or whenever changes to the record are reported, and examines: 1) the portion of incoming transactions that fail, 2) the portion of incoming transactions that are truly in error, and 3) the probability that the stored MIS record is in error for any reason. It is implicitly assumed that a piece of data (i.e., birth date, mother's maiden name, and rank) is accurate if it reflects the truth in the real world. A key result in this research is a mathematical relationship for the stored MIS record nonconformity rate as a function of the quality of the incoming data, the various processing times, the likelihood of Type I and II errors, and the probability distribution for the inter-transaction time. It is shown how the mathematical result can be used to forecast the level of improvement in the accuracy of the MIS record if corrective actions are taken.

Focusing on verification processes and building upon Morey's work, Paradice and Fuerst [88] develop a verification

TABLE V
DATA QUALITY LITERATURE RELATED TO DISTRIBUTION

Section	IV.A		IV.B			IV.C		IV.D	IV.E	IV.F	IV.G
Research		IV.B.1	IV.B.2	IV.B.3	IV.C.1	IV.C.2	IV.C.3				
[McCarthy, 1982]									✓		
[Page & Kaomea, 1994]						✓		✓	✓		
[Wang, Reddy & Gupta, 1993]						✓	✓	✓	✓		

mechanism based on statistical classification theory [61], whereby MIS records are classified as being either "correct" or "erroneous." Acceptance as a "correct" record by the verification mechanism occurs when the mechanism determines that the record is more similar to correct records, whereas rejection as an "erroneous" record occurs when the mechanism determines that the record is more similar to erroneous records. They use this as a benchmark for comparing actual error rates with the theoretically smallest attainable error rates, and suggest a method for assessing an organization's data quality. They also offer guidelines for deriving values for parameters of their data quality model.

Janson [58] demonstrates that exploratory statistical techniques [109] can significantly improve data quality during all phases of data validation. It is argued that data validation, if it is to be done successfully, requires knowledge of the underlying data structures. This is particularly crucial when the data are collected without any prior involvement by the analyst. Exploratory statistical techniques are well suited to the data validation effort because they can aid in identifying cases with data items that are suspect and likely to be erroneous, and in exploring the existence of functional relationships or patterns that can provide a basis for data cleaning.

Strong [100] examines how expert systems can be used to enhance data quality by inspecting and correcting non-conforming data items. A model is developed to evaluate the quality of the data products produced from a data production process [101]. It is also used to evaluate the process itself, and thus provides information to the research and development role for improving the data production process [103]. The process analysis indicates that it can be difficult to determine the quality of raw data inputs, data processing, and data products because what may be viewed as data and process irregularities by management may be viewed as necessary data production process flexibility by producers and consumers of data.

With some similarities to the above work, there are a number of studies that focus on the data editing function of survey data and the changing of data when errors are detected [41] [44], [59], [70], [73], [79]. These studies address data verification and correction through anomalies resolution and missing-values imputation in questionnaire data prior to processing. McKeown [79], for example, establishes probabilities that selected data fields were correct and contends that "data editing and imputation were separate and distinct." Garfinkel et al. [44] use experts to establish feasibility constraints, and develop algorithms that are "particularly relevant to problems in which data of every record are indispensable."

Oman and Ayers [86] provide a feedback loop for the correction of non-conforming data items. The method tabulates the volume of data reported, counts the number of errors, and divides the number of correct data by the total volume of data to produce a statistic on "percent correct" which is the "bottom line" statistic to be used in scoring data quality, and in providing feedback to reporting organizations. The analysis of the measurement shows marked improvement in the first half year of the effort and slow but steady progress overall.

The data tracking technique [52], [89], [92] is another verification mechanism of the quality of data in the data-product manufacturing process. A combination of statistical control and the manual identification of errors and their sources is used to systematically locate the cause of errors, and therefore, offers a framework for production verification. The data tracking technique in its present form, focuses primarily on certain data items and their consistency at different stages of their life cycle, and involves a significant amount of man-machine interaction.

The attribute-based approach [57], [87], [115], [116], [117], [119] can also be applied to verify the quality of data. By examining data quality indicators (such as source, time, and procurement method) or data quality parameters (such as timeliness, credibility, and completeness), the producer can verify, at different stages of the data product manufacturing process, whether the data conform with the specified requirements. When non-conforming data are identified, their data quality indicators can be used to trace back to the source of the problem so that appropriate corrective action may be taken.

Finally, as mentioned earlier, Svanks [104] reports on the actual development of an integrity analysis system that consists of seven steps. These steps, in some sense, cover the overall quality aspects in data production. Work on auditing can also be applied to the verification process.

D.1. Analysis of Research.

A body of research exists that can be related to the production of data products. Still, much more research is needed because there is a significant gap between the current state-of-the-art in data product production and the level of maturity required to establish a set of guidelines and standards in the data quality area similar to those established in ISO9000. Possible future research directions include: 1) developing "standard" data quality metrics or acceptance tests that could be used at the design stage or during the production process, 2) establishing criteria for the selection of qualified data suppliers, 3) developing a mechanism or tool to manage data quality shelf-life (out-of-date data) and deterioration control (data corruption), 4) studying the process, data flow and hu-

TABLE VI
DATA QUALITY LITERATURE RELATED TO PERSONNEL MANAGEMENT

Section	IV.A		IV.B			IV.C		IV.D	IV.E	IV.F	IV.G
Research		IV.B.1	IV.B.2	IV.B.3	IV.C.1	IV.C.2	IV.C.3				
[Maxwell, 1989]										✓	✓
[Oman & Ayers, 1988]	✓							✓		✓	
[Spirig, 1987]										✓	
[Te'eni, 1993]										✓	

man or automated procedures, 5) developing a method for positive identification and control of all non-conforming material (poor quality data), and 6) investigating the link between data quality problems and revision of the procedures used to detect and eliminate problems.

E. Distribution

Data quality research that can be related to distribution deals with data annotation and encapsulation, as summarized in Table V.

McCarthy [77] proposes self-describing data files and a metadata management approach at the schema level to annotate the data contained in the file. A similar annotation could be used to package data products. This would indicate, for example, what the data product is (identification), how it should be installed, etc. A direct analogy to physical products can easily be seen. For physical products, labels are required to describe product features such as expiration date (for food and drugs) or safety information (for electronic components). In the object-oriented field, an important concept is *encapsulation* by which data and procedures are "packaged" together. Some preliminary research has been devoted to apply the encapsulation concept to the packaging of data products [87], [116].

E.1. Analysis of Research.

Researchers could examine the literature in the area of physical product distribution to identify how that body of knowledge can be adapted to data product distribution. For instance, establishing an analogy could be a research goal. Consider data marting, an emerging concept in the database industry. Data marts can be set up through features such as a *data pipeline* that can be found in many commercial software packages. However, data products are not as tangible as physical products, and additional copies of data products can be produced at almost negligible cost when compared to physical products. Thus, adapting the knowledge of physical product distribution to data product distribution may not be a straightforward task. A thorough examination of data quality documentation is also required to assess need, role, and implementation strategies. In addition, researchers need to establish a customer feedback system for data quality, and to accurately define customer requirements.

The ultimate research goal in this area is to ensure that a data product delivered to data consumers meets the quality requirements specified for it. Finally, research could be pursued to extend the capabilities of current database management systems to handle identification, packaging, installation, delivery, and after-sales servicing of data products, as well as documentation and records for data products. We note in passing that data products are not necessarily "distributed"; but rather are made "available" or "accessible" to the user. They are stored in databases (data warehouses) where users can access and retrieve the relevant portion when needed. Thus, data accessibility can be interpreted as relevant to data product distribution.

F. Personnel Management

There have only been a few attempts by researchers to either address or analyze personnel issues within the context of data quality. These are summarized in Table VI.

Te'eni [107] proposes a general framework for understanding data production that incorporates the person-environment fit and the effect of an employee's ability and motivation. The research concentrates on the people who produce the data, examining the processes they employ and how the processes lead to the production of high quality data. The belief is that some of the problems in producing effective data are more likely to occur when one worker creates data and another uses it; that is, when data production is separated from data use. Data production problems are postulated to arise when there is a poor fit between the data producers' needs and resources, and the organization's demands and resources.

Maxwell [75] recognizes the need to improve the quality of data in human information systems databases as an important personnel issue. Many well-known examples of poor quality data in personnel databases are often cited; for example, a person's name may appear as "F. Wang" in one entry of a database and as "Forea L. Wang" in another. "Base salary" might include overtime premiums for one set of employees, but not for another. Maxwell proposes that, to improve human resource data quality, three issues need to be examined: 1) data ownership and origination, 2) accuracy-level requirements of specific data, and 3) procedural problems correctable through training and improved communications. A similar observation on data ownership has been made by Spirig [98]. He addresses some of the issues involved in interfacing payroll and personnel systems and suggests that, when data ownership becomes separated from the data originator, no system can retain data quality for very long.

Finally, Oman and Ayers [86] report on a case study of how a company's employees identified the need to improve data quality and brought it to management's attention. This organizational awareness resulted in action, by both the employees involved and by top management, to raise the level of data

TABLE VII
DATA QUALITY LITERATURE RELATED TO LEGAL FUNCTION

Section	IV.A		IV.B			IV.C		IV.D	IV.E	IV.F	IV.G
Research		IV.B.1	IV.B.2	IV.B.3	IV.C.1	IV.C.2	IV.C.3				
[Laudon, 1986]											✓
[Maxwell, 1989]										✓	✓
[Wright, 1992]											✓

quality in a large MIS database that obtained its raw data from over twenty subordinate reporting organizations. Data quality was improved by users searching for other sources of information and by using competing systems.

F.1. Analysis of Research.

There is some recognition of the awareness and motivational issues involved in obtaining high quality data. We do not see an immediate need for research in this area as it is likely that existing TQM techniques for personnel could be applied. Furthermore, it is more important to address management responsibilities first. After top management develops a data quality policy, personnel management issues related to data quality can be addressed. In the long term, research on personnel could include the following:

1) The development of an incentive plan to motivate employees to strive for high data quality. Case studies could be used to gain insights into what kinds of incentive structures are appropriate.
2) The development of a set of measures that could be used by the organization to monitor the quality level of the data obtained and generated by employees, and the creation of a feedback mechanism for the employees. For example, if one department is evaluated by the number of transactions that occur and another by daily balances, these discrepancies need to be understood.
3) The analysis of existing successful compensation and reward structures in various types of companies in order to understand how they work, so that successful approaches could be adapted to data quality.

G. Legal Function

Within the context of the framework, the legal issues surrounding data products include enhancing data product safety and minimizing product liability. The research efforts in this area are summarized in Table VII.

Laudon [69] examines the quality of data in the criminal-record system of the United States (a large inter-organizational system). Samples of records that deal with criminal history and warrant records are compared to hard-copy original documents of known quality. The results show that a large number of people could be at risk of being falsely detained, thus affecting their constitutional rights. Maxwell [75] identifies legal requirements as an important reason for ensuring the accuracy of employee-originated data. For example, Section 89 of the Internal Revenue Code requires that, unless 80% of a company's nonhighly compensated employees are covered by a given plan, then the plan must pass a series of tests on eligibility. This, in turn, places a priority on the need for accuracy of employee data. From the legal viewpoint, the data products produced by human resources information systems must be accurate, otherwise the company would be breaking the law. Wright [121], working in the electronic data interchange (EDI) area, examines the need to legally prove the origin and content of an electronic message.

Privacy and security are two other areas related to legal function. However, both are well-established fields and therefore, will not be discussed further.

G.1. Analysis of Research.

No work has been found on safety or liability limitations of data products. It is evident that data products are increasingly available in the (electronic) market. Some of them are produced by a single information provider (e.g., mailing lists, value-added information services, etc.) and others are produced through inter-organizational information systems that involve more than one legal entity. The legal ramifications of data products will become increasingly important given the trend toward information highways. The Commission of the European Community published a proposal for a Council Directive concerning the protection of individuals in relation to the processing of personal data [24]. The Directive contains chapters on the rights of data subjects, data quality, and provisions relating to liability and sanctions. From the data product liability viewpoint, research needs to be pursued that would investigate the effect on the parties involved of failed data products produced by an inter-organizational system. This includes everyone from the suppliers of the raw data to the parties who will either be using the data product directly or be affected by it. In addition, there is a requirement for methods to be developed on how to avoid liability issues.

V. CONCLUDING REMARKS

The database literature refers to data quality management as ensuring 1) syntactic correctness (e.g., constraints enforcement that prevents "garbage data" from being entered into the database), and 2) semantic correctness (data in the database truthfully reflect the real world situation). This traditional approach of data quality management leads to techniques such as integrity constraints, schema integration, and concurrency control. Although critical to data quality, these techniques fail to address some issues that are important to users. Many databases are plagued with erroneous data or data that do not meet users' needs. This approach fails to incorporate much of the results from the literature we have reviewed.

To overcome these problems, we took a practitioner's perspective and developed a framework for identifying and studying data quality issues. This framework consists of seven elements: management responsibilities, operation and assurance costs, research and development, production, distribution, personnel management, and legal function.

The principle findings of our analysis of the data quality literature, based on this framework, are as follows:

- First, there is a clear need to develop techniques that help management deliver quality data products. These techniques include quality policies and data quality systems.
- Second, the costs of external data failure and the complementary costs of data quality assurance need to be evaluated.
- Third, there is a need to study the link between poor data quality and procedures to detect and eliminate problems.
- Fourth, there are fundamental technical needs for an overall data quality metric and for a way to express rigorous quality requirements for a data product design.

This framework has proven to be effective in recognizing and organizing the literature in data quality management. Such a framework provides a vocabulary for discussing the various aspects of data quality that organizations increasingly experience. The framework has helped identify directions where research should be conducted if the ultimate goal of organizational information systems is to serve the needs of their users.

Acknowledgments

Work reported herein has been supported, in part, by MIT's Total Data Quality Management (TDQM) Research Program, the Univ. of Rochester's William E. Simon Graduate School of Business Administration, MIT's International Financial Services Research Center (IFSRC), the Naval Command, Control and Ocean Surveillance Center under Tactical Image Exploitation (TIX) contract N66001-92-D-0092, and Advanced Research Projects Agency and USAF/Rome Laboratory, under USAF contract F30602-93-C-0160.

Thanks are due to Peter Chen, Yang Lee, Stuart Madnick, and Diane Strong for their helpful comments on earlier versions of this paper. Thanks are also due to Nancy Chin and Terry Wong for their assistance in preparing this manuscript.

References

[1] N. Agmon and N. Ahituv, "Assessing data reliability in an information system," *J. Management Information Systems*, vol. 4, no. 2, pp. 34-44, 1987.

[2] N. Ahituv, "A systematic approach toward assessing the value of an information system," *MIS Quarterly*, vol. 4, no. 4, pp. 61-75, 1980.

[3] T. Amer, A.D. Bailey, and P. De, "A review of the computer information systems research related to accounting and auditing," *J. Information Systems*, vol. 2, no. 1, pp. 3-28, 1987.

[4] S.E. Arnold, "Information manufacturing: The road to database quality," *Database*, vol. 15, no. 5, pp. 32, 1992.

[5] J.E. Bailey and S.W. Pearson, "Development of a tool for measuring and analyzing computer user satisfaction," *Management Science*, vol. 29, no. 5, pp. 530-545, 1983.

[6] R. Bailey, *Human Error Computer Systems*. Englewood Cliffs, N.J.: Prentice Hall, 1983.

[7] D.P. Ballou and H.L. Pazer, "The impact of inspector fallibility on the inspection policy serial production system," *Management Science*, vol. 28, no. 4, pp. 387-399, 1982.

[8] D.P. Ballou and H.L. Pazer, "Modeling data and process quality multi-input, multi-output information systems," *Management Science*, vol. 31, no. 2, pp. 150-162, 1985.

[9] D.P. Ballou and H.L. Pazer, "Cost/quality tradeoffs for control procedures information systems," *OMEGA: Int'l J. Management Science*, vol. 15, no. 6, pp. 509-521, 1987.

[10] D.P. Ballou and H.L. Pazer, "A framework for the analysis of error conjunctive, multi-criteria, satisficing decision processes," *J. Decision Sciences Inst.*, vol. 21, no. 4, pp. 752-770, 1990.

[11] D.P. Ballou and H.L. Pazer, "Designing information systems to optimize the accuracy-timeliness tradeoff," *Information Systems Research* (forthcoming), 1995.

[12] D.P. Ballou and K.G. Tayi, "Methodology for allocating resources for data quality enhancement," *Comm. ACM*, vol. 32, no. 3, pp. 320-329, 1989.

[13] D.P. Ballou, R.Y. Wang, H. Pazer, and K.G. Tayi, *Modeling Data Manufacturing Systems To Determine Data Product Quality*, (No. TDQM-93-09). Cambridge, Mass.: Total Data Quality Management Research Program, MIT Sloan School of Management, 1993.

[14] C. Batini, M. Lenzirini, and S. Navathe, "A comparative analysis of methodologies for database schema integration," *ACM Computing Surveys*, vol. 18, no. 4, pp. 323-364, 1986.

[15] P.A. Bernstein and N. Goodman, "Concurrency control distributed database systems," *Computing Surveys*, vol. 13, no. 2, pp. 185-221, 1981.

[16] B. Blaylock and L. Rees, "Cognitive style and the usefulness of information," *Decision Sciences*, vol. 15, no. 1, pp. 74-91, 1984.

[17] G. Bodnar, "Reliability modeling of internal control systems," *Accounting Rev.*, vol. 50, no. 4, pp. 747-757, 1975.

[18] P. Bowen, "Managing data quality accounting information systems: A stochastic clearing system approach," unpublished PhD dissertation, Univ. of Tennessee, 1993.

[19] M.L. Brodie, "Data quality information systems, information, and management," vol. 3, pp. 245-258, 1980.

[20] W. Bulkeley, "Databases are plagued by reign of error," *Wall Street J.*, p. B6, May 26,1992.

[21] M. Bunge, *Ontology I: The Furniture of the World*, Treaties on Basic Philosophy, vol. 3. Boston, Mass.: D. Reidel Publishing, 1977.

[22] M. Bunge, *Ontology II: A World of Systems*. Treaties on Basic Philosophy, vol. 4. Boston, Mass.: D. Reidel Publishing, 1979.

[23] D. Burns and J. Loebbecke, "Internal control evaluation: How the computer can help," *J. Accountancy*, vol. 140, no. 2, pp. 60-70, 1975.

[24] S. Chalton, "The draft directive on data protection: an overview and progress to date," *Int'l Computer Law Adviser*, vol. 6, no. 1, pp. 6-12, 1991.

[25] P.P. Chen, "The entity-relationship model–Toward a unified view of data," *ACM Trans. Database Systems*, vol. 1, pp. 166-193, 1976.

[26] P.S. Chen, *The Entity-Relationship Approach, Information Technology Action: Trends and Perspectives*, R.Y. Wang, ed. Englewood Cliffs, N.J.: Prentice Hall, 1993.

[27] E.F. Codd, "A relational model of data for large shared data banks," *Comm. ACM*, vol. 13, no. 6, pp. 377-387, 1970.

[28] E.F. Codd, "Relational database: A practical foundation for productiv-

ity," 1981 ACM Turing Award Lecture, *Comm. ACM*, vol. 25, no. 2, pp. 109-117, 1982.

[29] R.B. Cooper, "Decision production–A step toward a theory of managerial information requirements," *Proc. Fourth Int'l Conf. on Information Systems*, pp. 215-268, Houston, Tex.,, 1983.

[30] P. Cronin, "Close the data quality gap through total data quality management," *MIT Management*, June 1993.

[31] P.B. Crosby, *Quality is Free*. New York: McGraw-Hill, 1979.

[32] P.B. Crosby, *Quality Without Tears*. New York: McGraw-Hill, 1984.

[33] B.E. Cushing, "A mathematical approach to the analysis and design of internal control systems," *Accounting Rev.*, vol. 49, no. 1, pp. 24-41, 1974.

[34] C.J. Date, *An Introduction to Database Systems*, Fifth edition. Reading, Mass.: Addison-Wesley, 1990.

[35] G.P.A. Delen and B.B. Rijsenbrij, "The specification, engineering, and measurement of information systems quality," *J. Systems Software*, vol. 17, no. 3, pp. 205-217, 1992.

[36] W.H. Delone and E.R. McLean, "Information systems success: The quest for the dependent variable," *Information Systems Research*, vol. 3, no. 1, pp. 60-95, 1992.

[37] E.W. Deming, *Out of the Crisis*. Cambridge, Mass.: MIT Center for Advanced Eng. Study, 1986.

[38] D.E. Denning and P.J. Denning, "Data Security," *ACM Computing Surveys*, vol. 11, no. 3, pp. 227-250, 1979.

[39] J.C. Emery, *Organizational planning and control systems: Theory and technology*. New York: Macmillan, 1969.

[40] A.V. Feigenbaum, *Total Quality Control*, Third edition. New York: McGraw-Hill, 1991.

[41] I.P. Fellegi and D. Holt, "A systematic approach to automatic edit and imputation," *J. Am. Statistical Assoc.*, vol. 71, no. 353, pp. 17-35, 1976.

[42] G. Feltham, "The value of information," *Accounting Rev.*, vol. 43, no. 4, pp. 684-696, 1968.

[43] K.T. Fields, H. Sami, and G.E. Sumners, "Quantification of the auditor's evaluation of internal control data base systems," *J. Information Systems*, vol. 1, no. 1, pp. 24-77, 1986.

[44] R.S. Garfinkel, A.S. Kunnathur, and G.E. Liepens, "Optimal imputation of erroneous data: Categorical data, general edits," *Operations Research*, vol. 34, no. 5, pp. 744-751, 1986.

[45] Gartner, "Data pollution can choke business process reengineering," Gartner Group Inside Industry Services, pp. 1, 1993.

[46] D.A. Garvin, "Quality on the line," *Harvard Business Rev.*, vol. 61, no. 5, pp. 65-75, 1983.

[47] S.M. Groomer and U.S. Murthy, "Continuous auditing of database applications: An embedded audit module approach," *J. Information Systems*, vol. 3, no. 2, pp. 53-69, 1989.

[48] D. Halloran et al., "Systems development quality control," *MIS Quarterly*, vol. 2, no. 4, pp. 1-12, 1978.

[49] S. Hamilton and N. Chervany, "Evaluating information system effectiveness–Part I: Comparing evaluation approaches," *MIS Quarterly*, vol. 5, no. 3, pp. 55-69, 1981.

[50] S.S. Hamlen, "A chance constrained mixed integer programming model for internal control design," *Accounting Rev.*, vol. 55, no. 4, pp. 578-593, 1980.

[51] J.V. Hansen, "Audit considerations distributed processing systems," *Comm. ACM*, vol. 26, no. 5, pp. 562-569, 1983.

[52] Y.U. Huh, F.R. Keller, T.C. Redman, and A.R. Watkins, "Data Quality," *Information and Software Technology*, vol. 32, no. 8, pp. 559-565, 1990.

[53] J. Iivari and E. Koskela, "The PIOCO model for information systems design," *MIS Quarterly*, vol. 11, no. 3, pp. 401-419, 1987.

[54] ISO, *ISO9000 Int'l Standards for Quality Management*. Geneva: Int'l Organization for Standards, 1992.

[55] B. Ives and M. Olson, "User involvement and MIS success: A review of research," *Management Science*, vol. 30, no. 5, pp. 586-603, 1984.

[56] B. Ives, M.H. Olson, and J.J. Baroudi, "The measurement of user information satisfaction," *Comm. ACM*, vol. 26, no. 10, pp. 785-793, 1983.

[57] Y. Jang, A.T. Ishii, and R.Y. Wang, "A qualitative approach to automatic data quality judgment," *J. Organizational Computing* (forthcoming), 1995.

[58] M. Janson, "Data quality: The Achilles heel of end-user computing," *Omega J. Management Science*, vol. 16, no. 5, pp. 491-502, 1988.

[59] M.A. Jaro, "Current record linkage research," *Proc. Am. Statistical Assoc.*, pp. 140-143, 1985.

[60] J.R. Johnson, R.A. Leitch, and J. Neter, "Characteristics of errors accounts receivable and inventory audits," *Accounting Rev.*, vol. 56, no. 2, pp. 270-293, 1981.

[61] R.A. Johnson and D.W. Wichern, *Applied Multivariate Statistical Analysis*, Second edition. Englewood Cliffs, N.J.: Prentice Hall, 1988.

[62] J.W. Jones and R. McLeod Jr., "The structure of executive information systems: An exploratory analysis," *Decision Sciences*, vol. 17, pp. 220-249, 1986.

[63] J.M. Juran, *Juran on Quality by Design: The New Steps for Planning Quality into Goods and Services*. New York: Free Press,, 1992.

[64] K.K. Kim, "User satisfaction: A synthesis of three different perspectives," *J. Information Systems*, vol. 4, no. 1, pp. 1-12, 1989.

[65] W. King and B.J. Epstein, "Assessing information system value: An experiment study," *Decision Sciences*, vol. 14, no. 1, pp. 34-45, 1983.

[66] C.H. Kriebel, "Evaluating the quality of information systems," *Design, and Implementation of Computer Based Information Systems*, N. Szysperski and E. Grochla, eds. Germantown: Sijthtoff and Noordhoff, 1979.

[67] A. Kumar and A. Segev, "Cost and availability tradeoffs replicated data concurrency control," *ACM Trans. Database Systems*, vol. 18, no. 1, pp. 102-131, 1993.

[68] D.F. Larcker and V.P. Lessig, "Perceived usefulness of information: A psychological examination," *Decision Sciences*, vol. 11, no. 1, pp. 121-134, 1980.

[69] K.C. Laudon, "Data quality and due process large interorganizational record systems," *Comm. ACM*, vol. 29, no. 1, pp. 4-11, 1986.

[70] G.E. Liepens, R.S. Garfinkel, and A.S. Kunnathur, "Error localization for erroneous data: A survey," *TIMS/Studies the Management Science*, vol. 19: pp. 205-219, 1982.

[71] G.E. Liepins, "Sound data are a sound investment," *Quality Progress*, vol. 22, no. 9, pp. 61-64, 1989.

[72] G.E. Liepins and V.R.R. Uppuluri, eds., *Data Quality Control: Theory and Pragmatics*. D.B. Owen, vol. 112. New York: Marcel Dekker, 1990.

[73] R.J.A. Little and P.J. Smith, "Editing and imputation for quantitative survey data," *J. Am. Statistical Assoc.*, vol. 82, no. 397, pp. 56-68, 1987.

[74] G.E. Liepins and V.R.R. Uppuluri, *Accuracy and Relevance and the Quality of Data*, A.S. Loebl, ed., vol. 112. New York: Marcel Dekker, 1990.

[75] B.S. Maxwell, "Beyond 'data validity': Improving the quality of HRIS data," *Personnel*, vol. 66, no. 4, pp. 48-58, 1989.

[76] J.A. McCall, P.K. Richards, and G.F. Walters, *Factors Software Quality*, (No. F030602-76-C-0417). Electronic Systems Division and Rome Air Development Center, 1977.

[77] J.L. McCarthy, "Metadata management for large statistical databases," *Proc. Eighth Int'l Conf. on Very Large Databases*, pp. 234-243, Mexico City, 1982.

[78] A.M. McGee, *Total Data Quality Management, Zero Defect Data Capture*, (No. TDQM-92-07). Cambridge, Mass.: Total Data Quality Management Research Program, MIT Sloan School of Management, 1992.

[79] P.G. McKeown, "Editing of continuous survey data," *SIAM J. Scientific and Statistical Computing*, pp. 784-797, 1984.

[80] N. Melone, "A theoretical assessment of the user-satisfaction construct information systems research," *Management Science*, vol. 36, no. 1, pp. 598-613, 1990.

[81] H. Mendelson and A. Saharia, "Incomplete information costs and database design," *ACM Trans. Database Systems*, vol. 11, no. 2, pp. 159-185, 1986.

[82] J. Miller and B.A. Doyle, "Measuring the effectiveness of computer-based information systems the financial services sector," *MIS Quarterly*, vol. 11, no. 1, pp. 107-124, 1987.

[83] K.I.J. Mollema, "Quality of information and EDP audit," *Informatie*, vol. 33, nos. 7-8, pp. 482-485, 1991.

[84] R.C. Morey, "Estimating and improving the quality of information the MIS," *Comm. ACM*, vol. 25, no. 5, pp. 337-342, 1982.

[85] D.R. Nichols, "A Model of auditor's preliminary evaluations of internal control from audit data," *The Accounting Rev.*, vol. 62, no. 1, pp. 183-190, 1987.

[86] R.C. Oman and T.B. Ayers, "Improving data quality," *J. Systems Management*, vol. 39, no. 5, pp. 31-35, 1988.

[87] W. Page and P. Kaomea, "Using quality attributes to produce optimal tactical information," *Proc. Fourth Ann. Workshop on Information Technologies and Systems*, pp. 145-154, Vancouver, B.C., Canada, 1994.

[88] D.B. Paradice and W.L. Fuerst, "An MIS data quality methodology based on optimal error detection," *J. Information Systems*, vol. 5, no. 1, pp. 48-66, 1991.

[89] R.W. Pautke and T.C. Redman, "Techniques to control and improve quality of data large databases," *Proc. of Statistics Canada Symp. 90*, pp. 319-333, Canada, 1990.

[90] M. Porter and V.E. Millar, "How information gives you competitive advantages," *Harvard Business Rev.*, vol. 63, no. 4, pp. 149-160, 1985.

[91] M.E. Porter, *Competitive Advantage*. New York: Free Press, 1985.

[92] T.C. Redman, *Data Quality: Management and Technology*. New York: Bantam Books, 1992.

[93] T.C. Redman, "Improve data quality for competitive advantage," *Sloan Management Rev.*, vol. 36, no. 2, pp. 99-109, 1995.

[94] B. Ronen and I. Spiegler, "Information as inventory: A new conceptual view," *Information and Management*, vol. 21, pp. 239-247, 1991.

[95] J. Saraph, G. Benson, and R. Schroeder, "An instrument for measuring the critical factors for quality management," *Decision Sciences*, vol. 20, no. 4, pp. 810-829, 1989.

[96] N. Schneidewind, *Standard for a Software Quality Metrics Methodology*. Software Eng. Standards Subcommittee of the IEEE, 1990.

[97] J.C. Sparhawk Jr., "How does the Fed data garden grow? By deeply sowing the seeds of TQM.," *Government Computer News*, Jan. 18, 1993.

[98] J. Spirig, "Compensation: The up-front issues of payroll and HRIS interface," *Personnel J.*, vol. 66, no. 10, pp. 124-129, 1987.

[99] W.O. Stratton, "Accounting systems: The reliability approach to internal control evaluation," *Decision Sciences*, vol. 12, no. 1, pp. 51-67, 1981.

[100]D.M. Strong, "Design and evaluation of information handling processes," PhD dissertation, Carnegie Mellon Univ., 1988.

[101]D.M. Strong,. *Modeling Exception Handling and Quality Control Information Processes*, No. WP 92-36. Boston, Mass.: School of Management, Boston Univ., 1993.

[102]D.M. Strong, Y.W. Lee, and R.Y. Wang, *Beyond Accuracy: How Organizations are Redefining Data Quality*. (No. TDQM-94-07). Cambridge, Mass.: Total Data Quality Management (TDQM) Research Program, MIT Sloan School Of Management, 1994.

[103]D.M. Strong and S.M. Miller, "Exceptions and exception handling in computerized information processes," *ACM Trans. on Information Systems* (forthcoming), 1993.

[104] M.I. Svanks, "Integrity analysis: Methods for automating data quality assurance," *EDP Auditors Foundation*, vol. 30, no. 10, pp. 595-605, 1984.

[105]G. Taguchi, *Introduction to Off-line Quality Control*. Magaya, Japan: Central Japan Quality Control Assoc., 1979.

[106]A.U. Tansel et al., *Temporal Databases: Theory, Design, and Implementation*, S. Navathe, ed. Redwood City, Calif.: Benjamin/Cummings Publishing, 1993.

[107]D. Te'eni, "Behavioral aspects of data production and their impact on data quality," *J. Database Management*, vol. 4, no. 2, pp. 30-38, 1993.

[108]T.J. Teorey, D. Yang, and J.P. Fry, "A logical design methodology for relational databases using the extended entity-relationship model," *ACM Computing Surveys*, vol. 18, no. 2, pp. 197-222, 1986.

[109]J.W. Tukey, *Exploratory Data Analysis*. Reading, Mass.: Addison-Wesley, 1977.

[110]Y. Wand, "A proposal for a formal model of objects," *Object-Oriented Concepts, Databases, and Applications*, W. Kim and F. Lochovsky, eds. New York: ACM Press, 1989.

[111]Y. Wand and R.Y. Wang (1994), "Anchoring data quality dimensions ontological foundations," *Comm. ACM*, forthcoming.

[112]Y. Wand and R. Weber, "A model of control and audit procedure change evoloving data processing systems," *The Accounting Rev.*, vol. 64, no. 1, pp. 87-107, 1989.

[113] Y. Wand and R. Weber, "An ontological model of an information system," *IEEE Trans. Software Engineering*, vol. 16, no. 11, pp. 1,282-1,292, 1990.

[114]R.Y. Wang and H.B. Kon, "Towards total data quality management," *Information Technology Action: Trends and Perspectives*, R.Y. Wang, ed. Englewood Cliffs, N.J.: Prentice Hall, 1993.

[115]R.Y. Wang, H.B. Kon, and S.E. Madnick, "Data quality requirements analysis and modeling," *Proc. Ninth Int'l Conf. on Data Engineering*, pp. 670-677, Vienna, 1993.

[116]R.Y. Wang, M.P. Reddy, and A. Gupta, "An object-oriented implementation of quality data products," *Proc. Third Ann. Workshop Information Technologies and Systems*, pp. 48-56, Orlando, Fla., 1993.

[117] R.Y. Wang, M.P. Reddy, and H.B. Kon, "Toward quality data: An attribute-based approach," *J. Decision Support Systems*, (forthcoming), 1992.

[118]R.Y. Wang, D.M. Strong, and L.M. Guarascio, *Beyond Accuracy: What Data Quality Means to Data Consumers*, (No. TDQM-94-10). Cambridge, Mass.: Total Data Quality Management Research Program, MIT Sloan School of Management, 1994.

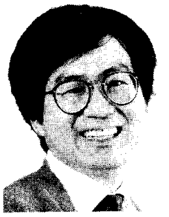
[119]Y.R. Wang and S.E. Madnick, "A Polygen model for heterogeneous database systems: The source tagging perspective," *Proc. 16th Int'l Conf. Very Large Databases*, pp. 519-538, Brisbane, Australia, 1990.

[120]R. Weber, *EDP Auditing: Conceptual Foundations and Practices*, second edition, McGraw-Hill Series MIS, G.B. Davis, ed. New York: McGraw-Hill, 1988.

[121]B. Wright, "Authenticating EDI: The case for internal record keeping," *EDI Forum*, pp. 82-84, 1992.

[122]S. Yu and J. Neter, "A stochastic model of the internal control system," *J. Accounting Research*, vol. 1, no. 3, pp. 273-295, 1973.

[123]R. Zmud, "Concepts, theories, and technlques: An empirical investigation of the dimensionality of the concept of information," *Decision Sciences*, vol. 9, no. 2, pp. 187-195, 1978.

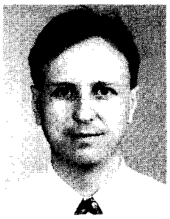


Richard Y. Wang received a PhD with a concentration in information technologies at the Massachusetts Institute of Technology Sloan School of Management, where he is now an associate professor of information technologies. He also serves as codirector of the Total Data Quality Management (TDQM) Research Program at MIT.

Prof. Wang has published extensively in the areas of data quality management, database management systems, and connectivity among information systems. In addition, he is the author of *Information Technologies: Trends and Perspectives* (Prentice Hall, 1993).

Veda C. Storey earned a BS (with distinction) from Mt. Allison University, New Brunswick, Canada, in 1978, an MBA from Queen's University, Ontario, Canada, in 1980, and a PhD in management information systems from the University of British Columbia, Canada, in 1986.

Dr. Storey is a member of the Computer and Information Systems Dept. at the William E. Simon Graduate School of Business Adminstration, University of Rochester, New York. Her research interests lie in database management systems and artificial intelligence. She has published papers in many leading academic journals and is the author of *View Creation: An Expert System for Database Design* (International Center for Information Technology Press, 1988), a book based on her doctoral dissertation.



Christospher P. Firth is the business architect for the Global Consumer Bank of Citibank Singapore. He is concerned with the strategic and tactical deployment of information technology within the bank, and has a special interest in the certification of large databases and data quality, and their effect on customer service.

Previously, Firth was with the Management of Technology Program at the MIT Sloan School of Management, where he was also a research affiliate in the MIT TDQM Research Program. He has extensive work experience in the financial services and the IT industry in Singapore, Hong Kong, France, and the U.K.