Diane M. Strong, Yang W. Lee, and Richard Y. Wang

Data Quality In Context

A new study reveals businesses are defining data quality with the consumer in mind.

Data-quality (DQ) problems are increasingly evident, particularly in organizational databases. Indeed, 50% to 80% of computerized criminal records in the U.S. were found to be inaccurate, incomplete, or ambiguous. The social and economic impact of poor-quality data costs billions of dollars. [5-7, 10].

Organizational databases, however, reside in the larger context of information systems (IS). Within this larger context, data is collected from multiple data sources and stored in databases. From this stored data, useful information[1] is generated for organizational decision-making.

[1]For consistency, we use the term "data" throughout this article to refer to both data and information. This avoids switching between terms as we switch between production and use of data.

DQ problems may arise anywhere in this larger IS context.[2] Thus, we argue for a conceptualization of data quality that includes this context.

Database research aims at ensuring the quality of data in databases. In the DQ area, existing research investigates DQ definitions [8, 11], modeling [1, 2], and control [6]. With few exceptions, however, DQ is treated as an intrinsic concept, independent of the context in which data is produced and used. This focus on intrinsic DQ problems in stored data fails to solve complex organizational problems. We attribute this failure, in part, to the lack of a broader DQ conceptualization. When quality problems are defined as errors in stored data, IS professionals may not recognize, and thus solve, the most critical DQ problems in organizations.

In contrast to this intrinsic view, it is well accepted in the quality literature that quality cannot be assessed independent of consumers who choose and use products [4]. Similarly, the quality of data cannot be assessed independent of the people who use data—data consumers. Data consumers' assessments of DQ are increasingly important because consumers now have more choices and control over their computing environment and the data they use. To solve organizational DQ problems, therefore, one must consider DQ beyond the intrinsic view. Moreover, one must move beyond stored data to include data in production and utilization processes.

Using qualitative analysis, we examined DQ projects from three leading-edge organizations and identified common patterns of quality problems. These patterns emerged because we used a broader conceptualization of DQ. Based on these patterns, we developed recommendations for IS professionals to improve DQ from the perspective of data consumers.

Definitions and Methods in Context

Production and storage of data has been conceptualized as a data manufacturing system [3, 9]. Central to this is the concept of a data production process that transforms data into information useful to data consumers. We identify three roles within data manufacturing systems: data producers (people, groups, or other sources who generate data); data custodians (people who provide and manage computing resources for storing and processing data); and data consumers (people or groups who use data). Each role is associated with a process or task: data producers are associated with data-production processes; data custodians with data storage, maintenance, and security; and data consumers with data-utilization processes, which may involve additional data aggregation and integration.

We define high-quality data as data that is fit for use by data consumers—a widely adopted criteria. This means that usefulness and usability are important aspects of quality. Using this definition, the characteristics of high-quality data (Table 1) consist of four categories: intrinsic, accessibility, contextual, and representational aspects. This data consumers' perspective is a broader conceptualization of DQ than the conventional intrinsic view.

Table 1. DQ categories and dimensions

DQ Category	DQ Dimensions
Intrinsic DQ	Accuracy, Objectivity, Believability, Reputation
Accessibility DQ	Accessibility, Access security
Contextual DQ	Relevancy, Value-Added, Timeliness, Completeness, Amount of data
Representational DQ	Interpretability, Ease of understanding, Concise representation, Consistent representation

We define a *DQ problem* as any difficulty encountered along one or more quality dimensions that renders data completely or largely unfit for use. We define a *DQ project* as organizational actions taken to address a DQ problem given some recognition of poor DQ by the organization. We intentionally include projects initiated for purposes other than improving DQ. For example, during conversion of data to a client/server system, poor DQ may be recognized and an improvement initiated.

To examine DQ problems in practice, we studied 42 DQ projects from three data-rich organizations: GoldenAir, an international airline; BetterCare, a hospital; and HyCare, a Health Maintenance Organization (HMO). In terms of industry position, attention to DQ, and information systems, these three firms are leaders, yet they exhibit sufficient variation for investigating data projects (Table 2). All have identified significant DQ problems, and are actively attending to them. This contrasts with many organizations that fail to address their quality problems.

This research employed qualitative data collection and analysis techniques. We collected data about these projects via interviews of data producers, cus-

[2]The term "information system" is sometimes used to mean a database or a computer system (including hardware and software). Our use of the phase "larger information systems context" covers the organizational processes, procedures, and roles employed in collecting, processing, distributing and using data.

Table 2. Site characteristics

Site Name* and Industry	Attention to DQ	IS Organization	Hardware and Software Environment
GoldenAir Airline	IS Development	IS is essentially a service bureau.	IBM-compatible mainframe with IMS databases and MMS.
BetterCare Hospital	DQ Administrator	Centralized IS organization reporting to finance VP in a centralized, functional firm.	PC-based client server environment with TRACE, a MUMPS-based database system.
HyCare HMO	Total Quality Management (TQM) Initatives	Powerful, centralized IS organization in a decentralized, divisional firm.	Heterogeneous hardware and software across divisions.

*All names are fictitious

todians, consumers, and managers. We organized each DQ project in terms of three problem-solving steps: *problem finding* (how the organization identified a DQ problem), *problem analysis* (what the organization determined the cause to be), and *problem resolution* that includes changing processes (changing the procedures for producing, storing, or using data) and changing data (updating the data value). Each project was analyzed using the DQ dimensions as content analysis codes. From the coded projects, we identified common patterns and sequences of dimensions attended to during DQ projects (Table 3).[3]

Table 3. DQ patterns in DQ projects

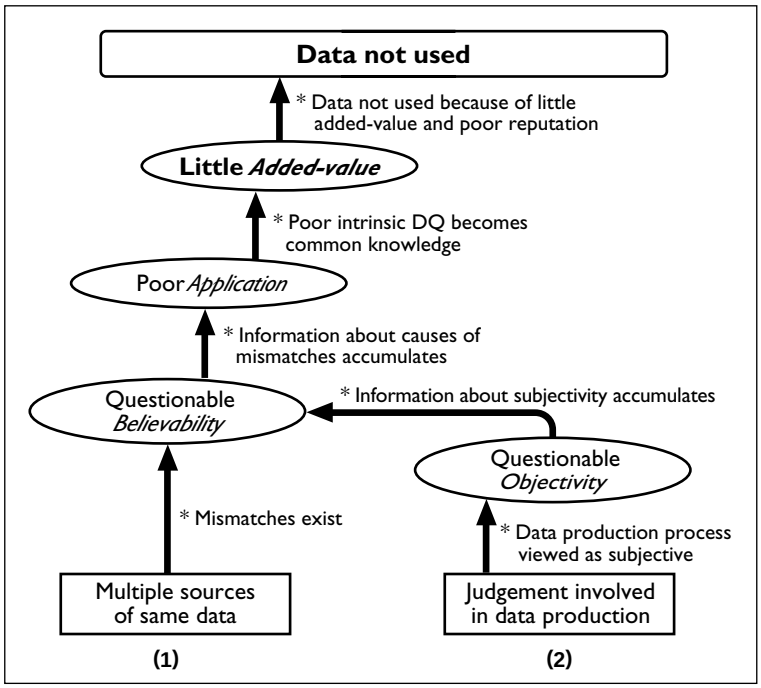
DQ Category	DQ Dimensions
Intrinsic DQ	Accuracy, Objectivity, Believability, Reputation
Accessibility DQ	Accessibility, Access security
Contextual DQ	Relevancy, Value-Added, Timeliness, Completeness, Amount of data
Representational DQ	Interpretability, Ease of understanding, Concise representation, Consistent representation

Intrinsic DQ Pattern

Mismatches among sources of the same data are a common cause of intrinsic DQ concerns. Initially, data consumers do not know the source to which quality problems should be attributed; they know only that data is conflicting. Thus, these concerns initially appear as *believability*[4] problems. Over time, information about the causes of mismatches accumulates from evaluations of the *accuracy* of different sources, which leads to a poor *reputation* for less accurate sources. (A reputation for poor quality can also develop with little factual basis.) As a reputation for poor-quality data becomes common knowledge, these data sources are viewed as having little *added value* for the organization, resulting in reduced use (Figure 1, subpattern 1).

Judgment or subjectivity in the data production process is another common cause (subpattern 2). For example, coded or interpreted data is considered to be of lower quality than raw, uninterpreted

Figure 1. Intrinsic DQ problem pattern



[3]An appendix containing method details and example projects is posted at http://web.mit.edu/tdqm.

[4]The italics signifies that believability is a DQ dimension. This convention will be used to highlight the interaction of DQ dimensions in a DQ project.

data. Initially, only those with knowledge of data production processes are aware of these potential problems, which appear as concerns about data *objectivity.* Over time, information about the subjective nature of data production accumulates, resulting in data of questionable believability and reputation and thus of little added value to data consumers. The overall result is reduced use of this suspect data.

Intrinsic DQ subpattern 1 was exhibited at all three research sites. GoldenAir has a history of mismatches between their inventory system data and physical warehouse counts. Warehouse counts serve as a standard against which to measure the accuracy of system data, for example, the system data source is

between internal HMO patient records and bills submitted by hospitals for reimbursement. For example, when the HMO is billed for coronary bypass surgery, the HMO patient record should indicate active, serious heart problems. Mismatches occur in both directions, hospital claims without HMO records of problems, and HMO records of problems without corresponding hospital claims. Initially, HyCare assumed the external (hospital) data was wrong; HMO staff perceived their data to be more believable and have a better reputation than those of hospitals. This general sense of the quality of sources, however, was not based on factual analysis.

Subpattern 2 occurred at both BetterCare and

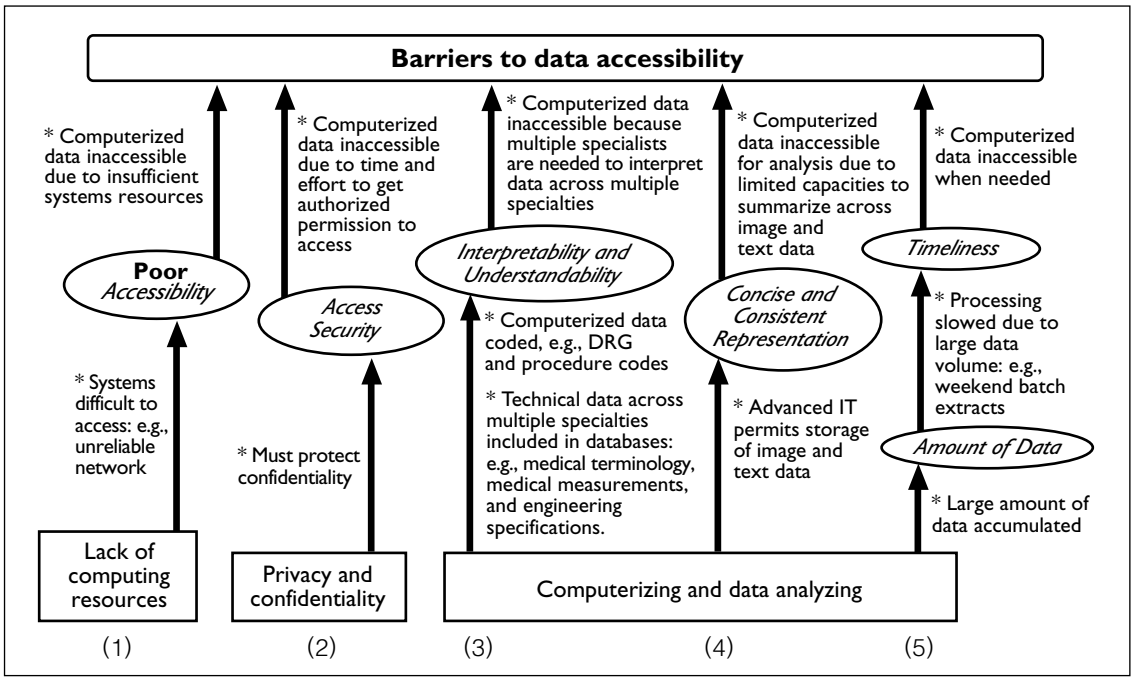


Figure 2. Accessibility DQ problem pattern

inaccurate and not believable, and is adjusted periodically to match actual warehouse counts. The system data gradually develops mismatches, however, and its reputation gradually worsens until the data is not used for decision-making.

At BetterCare, this subpattern occurred between TRACE[5] and STATUS.[6] Some data, like daily hospital bed utilization, is available from both systems. Nevertheless, it frequently have different values. Over time, TRACE has developed a reputation as an accurate source, and the use of STATUS has declined.

At HyCare, inconsistent data values occur

HyCare. Using doctors' and nurses' notes about patients, BetterCare's medical record coders designate diagnosis and procedure codes and corresponding diagnosis-related groups (DRG) codes for billing. Although coders are highly trained, some subjectivity remains. Thus, this data is considered to be less objective than raw data.

Data-production forms also contribute to reduced objectivity of data. At HyCare, doctors using preprinted forms with check boxes for specifying procedure codes generated a reduced range of procedures performed, as compared to doctors using free-form input. This variance affects the believability of this data.

The three organizations developed the following solutions for handling subpattern 1:

[5]TRACE is a database containing historical data extracted from the hospital's information and control system for use by managers making longer-term decisions and by medical researchers.

[6]STATUS is an operational system that records a snapshot of daily hospital resources.

- GoldenAir continues their cycle of physically counting inventory and adjusting system values whenever the mismatch becomes unacceptably large.
- BetterCare is rewriting STATUS. They are also designating single data production points for data items and improving computerized support for data production.
- HyCare's analysis of the causes of mismatches between hospital and internal data found problems with both sources. They fixed an edit check problem with their internal computer systems, fixed a data production problem in doctors' designation of active, serious problems for internal HMO records, and initiated joint DQ projects with associated hospitals.

These solutions manifest two different approaches to problem resolution: changing the systems or changing the production processes. GoldenAir focused on computer systems as the solution and ignored their data production processes. As a result, their processes continue to produce poor-quality data that increases data inaccuracies. In contrast, BetterCare's and HyCare's solutions involve both data production processes and computer systems, resulting in long-term DQ improvements.

BetterCare's efforts to designate single data production points deserve further discussion. Systems developed for different purposes sometimes require the same data, such as an indicator of patient severity in intensive care units in both STATUS and HICS. For HICS, a specialist examines the patient immediately before intensive care. For STATUS, an intensive-care nurse observes the patient during intensive care. These two observations can be different. To designate a single source, definitions and indicators of severity were agreed upon and both systems were changed to support this single data production source.

BetterCare's decision to rewrite STATUS illustrates reputation development. Like accounting systems that prohibit changes once the accounting period is closed, STATUS prohibits changes to the official daily record. STATUS's data is *consistent* across time, whereas TRACE's data is accurate because it is updated as needed. Although both systems are viewed as containing the "correct" data, TRACE developed a reputation as the system with high-quality data, whereas STATUS's data was considered to be suspect. As a result, STATUS is being rewritten with update routines.

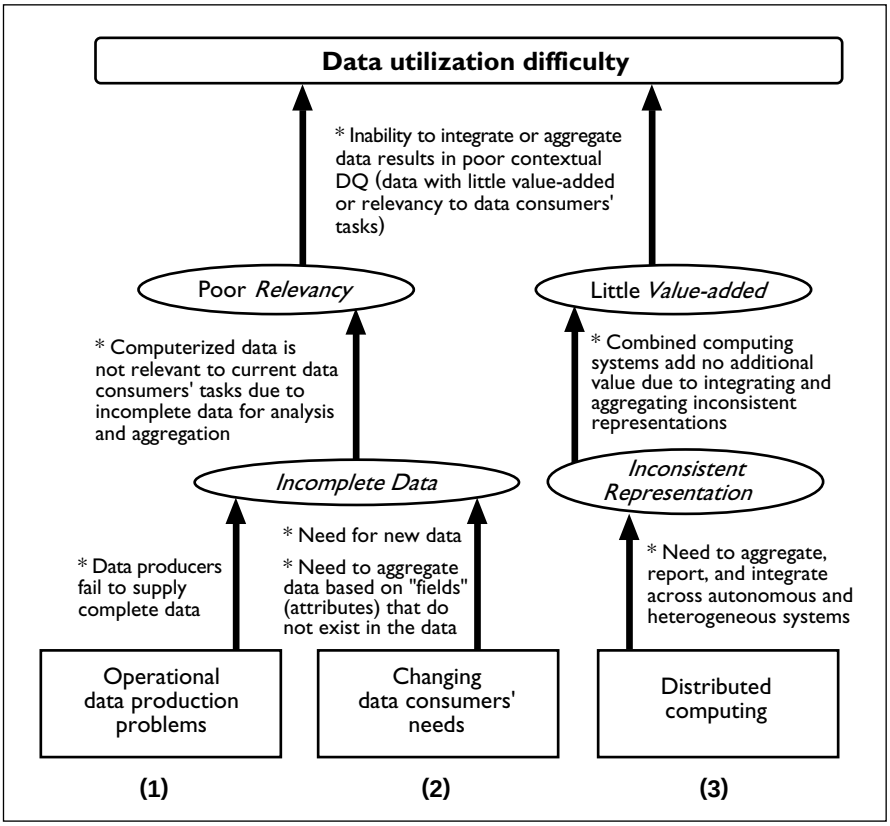


Figure 3. Contextual DQ problem pattern

Accessibility DQ Pattern

Accessibility DQ problems were characterized by underlying concerns about technical accessibility (Figure 2, subpatterns 1-2), data-representation issues interpreted by data consumers as accessibility problems (subpatterns 3-4), and data-volume issues interpreted as accessibility problems (subpattern 5).

GoldenAir provides a simple example of subpattern 1. When GoldenAir moved to its new airport, its computing operations remained at the old airport with *access* to data via unreliable data communications lines. Since reservations had priority, the unreliable lines resulted in inventory data accessibility problems. This, in turn, contributed to GoldenAir's inventory accuracy problems because updating took lower priority than other data-related tasks.

BetterCare had an accessibility DQ concern related to the confidential nature of patient records (subpattern 2). Data consumers realized the importance of access *security* for patient records, but they also perceived the permissions as barriers to accessibility. This, in turn, affects the overall reputation and *value* of this data. In addition, data custodians became barriers to accessibility because they could not provide data access without approval.

Subpattern 3 addresses concerns about *interpretability* and *understandability* of data. Coding systems for physician and hospital activities at BetterCare and HyCare are necessary for summarizing and grouping common diagnoses and procedures. The expertise required to interpret codes, however, becomes a barrier to accessibility; these codes are not understandable to most doctors and analysts. At HyCare, analyzing and interpreting across physician groups is a problem because they use different coding systems.

Medical data in text or image form also presents an interpretability problem (subpattern 4). Medical records include text written by doctors and nurses and images produced by medical equipment, such as X-rays and EKGs. This data is difficult to analyze across time for individual patients. Furthermore, analyzing trends across patients is difficult. Thus, data *representation* becomes a barrier to data accessibility. This data are inaccessible to data consumers because it is not in a representation that permits analysis.

Subpattern 5 addresses providing *relevant* data that adds value to tasks in a timely manner. For example, HyCare serves hundreds of thousands of patients resulting in several million patient records tracking medical history. Analyses of patient records usually require a weekend data extraction. In addition, companies purchasing HMO options are increasingly demanding evaluations of medical practices, resulting in an increased need for these analyses at HyCare. This pattern of a large *amount of data* leading to *timeliness* problems are interpreted as accessibility problems.

Subpattern 1 has straight-forward, though possibly costly, solutions. For example, GoldenAir is moving its computing facility to the new airport to avoid unreliable data communication lines. Subpattern 5 is also relatively easy to solve. For example, BetterCare's HICS generates 40GB of data per year. From this, TRACE extracts the most relevant data (totaling 5GB over 12 years) for historical and cross-patient analyses.

Subpatterns 3 and 4 are more difficult to solve. Although HyCare completely automated its medical records, including text and image data, to solve accessibility problems for individual patients, and problems with analyzing data across patients persist. At BetterCare, data consumers and custodians believe that an automated representation of text and image data would not solve their analyzability problems; thus, they partially automated their patient records.

Contextual DQ Pattern

We observed three underlying causes for data consumers' complaints that available data does not support their tasks: missing (*incomplete*) data, inadequately defined or measured data, data that could not be appropriately aggregated.

To solve these contextual DQ problems, specific projects were initiated to provide relevant data that adds value to the tasks of data consumers.

Subpattern 1 in Figure 3 addresses incomplete data due to operational problems. At GoldenAir, incomplete data in inventory transactions contributed to inventory data accuracy problems. For example, mechanics sometimes failed to record part numbers on their work activity forms. Because transaction data was incomplete, the inventory database could not be updated, which in turn produced inaccurate records. According to one supervisor, this was tolerated because "the primary job of mechanics is to service aircraft in a timely manner, not to fill out forms."

BetterCare's data was incomplete by design (subpattern 2), whereas GoldenAir's data was incomplete due to operational problems. By design, the amount of data in BetterCare's TRACE database is small enough to be accessible but complete enough to be relevant and add value to data consumers' tasks. As a result, data consumers occasionally complained about incomplete data.

Subpattern 3 addresses problems caused by integrating data across distributed systems. At HyCare, data consumers complained about inconsistent definitions and data representations across divisions, like DRG codes stored with decimal points in one division and without in another. Furthermore, basic utilization measures, such as hospital days per thousand patients, were defined differently across divisions. These problems were caused by autonomous design decisions in each division.

GoldenAir is considering bar code readers as data input mechanisms (subpattern 1). BetterCare's decision about the data to include in TRACE is reassessed as data consumers request additional data (subpattern 2), such as healthcare proxy and living will information were added.

This reassessment of TRACE data in the context of its relevance and value to data consumers goes beyond missing data. As healthcare reimbursement systems move from payments for procedures performed (fee for service) to payments for diagnosed diseases (prospective payment) to possibly payments for yearly care of patients (capitated payment), the basic unit of analysis for managerial decision-making in hospitals has changed from procedures, to hospital visits, to patients. When BetterCare tracked data by procedures, for example, they could answer questions about costs of blood tests, but not costs of treating heart attacks. Such analyses became necessary when hospital reimbursement changed to a fixed amount for treating each disease.

TRACE was developed in response to this anticipated change to prospective payments. Such a reimbursement system began in 1983 for Medicare. At that time, TRACE had the capability to aggregate across patient visits for similar diagnoses. Currently, the ability to aggregate across all in- and out-patient medical services delivered to each patient per year is being anticipated by BetterCare. Thus, TRACE is being extended with out-patient data and quality indicators because management anticipates these changes.

HyCare initiated DQ projects to develop common data definitions and representations for cross-divisional data (subpattern 3). The comprehensive data dictionary and corresponding data warehouse are their next steps.

Implications for IS Professionals

Our findings provide generalizable implications for IS professionals about solving intrinsic, accessibility, and contextual DQ problems.

Conventional DQ approaches employ control techniques (like edit checks, database integrity constraints, and program control of database updates) to ensure data quality. These approaches have improved intrinsic DQ substantially, especially the accuracy dimension. Attention to accuracy alone, however, does not correspond to data consumers' broader DQ concerns. Furthermore, controls on data storage are necessary but not sufficient. IS professionals also need to apply process-oriented techniques, like IS auditing [12], to the processes that produce this data.

Data consumers perceive any access barriers as accessibility problems. Conventional approaches treat accessibility as a technical, computer systems issue, not a DQ concern. That is, data custodians have provided access if data is technically accessible (such as when terminals and lines are connected and available, access permission is granted, and access methods are installed). To data consumers, however, accessibility goes beyond technical accessibility; it includes the ease with which they can manipulate this data to suit their needs.

These contrasting accessibility views are evident in our study. For example, advanced forms of data (medical image data) can now be stored as binary large objects (blobs). Although data custodians provide technical methods for accessing this new form of data, data consumers continued to experience this data as inaccessible. They need to analyze this data like they analyze traditional record-oriented data. Other examples of differing views of accessibility include

- Data combined across autonomous systems is technically accessible, but data consumers view it as inaccessible because similar data items are defined, measured, or represented differently.
- Coded medical data is technically accessible as text, but data consumers view it as inaccessible because they cannot interpret the codes.
- Large volumes of data is technically accessible, but data consumers view it as inaccessible because of excessive access time.

IS professionals must understand the difference between the technical accessibility they supply and the broad accessibility concerns of data consumers. Once this difference is clarified, technologies such as data warehouses can provide a smaller amount of more relevant data, and graphical interfaces can improve ease of access.

Data consumers evaluate DQ relative to their tasks. At any time, the same data may be needed for multiple tasks that require different quality characteristics. Furthermore, these quality characteristics will change over time as work requirements change. Therefore, providing high-quality data implies tracking an ever-moving target. Conventional approaches handle contextual DQ through techniques such as user requirements analysis and relational database query capabilities. They do not explicitly incorporate the changing nature of data consumers' task context.

Because data consumers perform many different tasks and the data requirements for these tasks change, contextual DQ means much more than good data requirements specification. Providing high-quality data along the dimensions of value and usefulness relative to data consumers' task contexts places a premium on designing flexible systems with data that can be easily aggregated and manipulated.

The alternative is constant maintenance of data and systems to meet changing data requirements.

Concluding Remarks

Existing research focuses on intrinsic aspects of DQ. It fails to address the broader concerns of data consumers. While intrinsic DQ aspects are important, organizations also initiate projects to address accessibility and contextual DQ issues. Accessibility DQ includes concerns about the ease of access and ease of understanding data. Contextual DQ includes concerns about how well data matches task contexts.

This research adopts a data-consumer perspective. The results confirm the importance of the quality categories and dimensions in our previous research [11]. They also enrich our understanding of how organizations experience DQ problems and which dimensions comprise these problems. For example, this research discovered that representational DQ dimensions are underlying causes of accessibility DQ problem patterns.

Some might argue our research findings can be attributed to poor management or poor IS organizations at our field sites. We reject such a claim. The organizations we studied are competent and address their DQ problems effectively. They are at the forefront of DQ practice. Others may agree with our findings, but argue that accessibility and contextual DQ fall outside the domain. We also reject such a view. To solve organizational DQ problems, IS professionals must attend to the entire range of concerns of data consumers.

The results of this research may be used as an empirical basis for building DQ theories about the nature of organizational DQ problems and their solutions. Given our results, new DQ theories will incorporate the task context of users and the processes by which users access and manipulate data to meet their task requirements. For example, a theory based on consumer marketing research could investigate when and how data consumers apply various DQ dimensions in choosing data for their tasks. Studies that focus on accessibility issues exemplify this approach.

The three patterns for how intrinsic, accessibility, and contextual DQ problems develop in organizations provides an empirical basis for studying organizational choices and actions about DQ improvement. For example, organizational theories can be applied to understand how organizations find and choose to solve DQ problems. Following a time-dependent decision processes perspective, solutions to DQ problems are found, implemented, learned, and improved, through adaptation over time. Following a perspective of organizational routines as sources of performance, TQM procedures and DQ administrators can establish organizational routines that improve DQ. Theories in information economics could also be applied to understanding organizational decisions about improvement.

In addition to theory building, studies of DQ solutions could use the DQ-problem patterns identified in this research as solution objectives. For example, known DQ problems will focus the search for organizational mechanisms that solve these problems. Finally, this research should be replicated in organizations such as financial firms, where data is their primary product. C

References

1. Ballou, D. P. and Pazer, H. L. Modeling data and process quality in multi-input, multi-output information systems. *Manage. Sci. 31*, 2 (1985), pp. 150–162.
2. Ballou, D. P. and Tayi, K. G. Methodology for allocating resources for data quality enhancement. *Commun. ACM 32*, 3 (1989), pp. 320–329.
3. Ballou, D. P., Wang, R. Y., Pazer, H., and Tayi, K. G. Modeling information manufacturing systems to determine information product quality. *Manage. Sci*. (accepted for publication, 1996).
4. Deming, E. W. *Out of the Crisis*. MIT Center for Advanced Engineering Study. Cambridge, Mass. 1986.
5. Laudon, K. C. Data quality and due process in large interorganizational record systems. *Commun. ACM 29*, 1 (1986), pp. 4–11.
6. Liepins, G. E. and Uppuluri, V. R. R., Eds. *Data Quality Control: Theory and Pragmatics*. D. B. Owen, (1990), Marcel Dekker, New York, N.Y.
7. Morey, R. C. Estimating and improving the quality of information in MIS. *Commun. ACM 25*, 5 (1982), pp. 337–342.
8. Wand, Y. and Wang, R. Y. Anchoring data quality dimensions in ontological foundations. *Commun. ACM 39*, 11 (1996), pp.86–95.
9. Wang, R. Y. and Kon, H. B. Towards Total Data Quality Management (TDQM). *Information Technology in Action: Trends and Perspectives*. R. Y. Wang, Ed. 1993. Prentice Hall, Englewood Cliffs, NJ.
10. Wang, R. Y., Storey, V. C. and Firth, C. P. A framework for analysis of data quality research. *IEEE Trans. Know. Data Eng. 7*, 4 (1995), pp. 623–640.
11. Wang, R. Y. and Strong, D. M. Beyond accuracy: What data quality means to data consumers. *J. Manage. Info. Syst. 12*, 4 (1996), pp. 5–34.
12. Weber, R. EDP *Auditing: Conceptual Foundations and Practices*. G. B. Davis, Ed. 1988. McGraw-Hill, New York, N.Y.

Diane M. Strong (dstrong@wpi.edu) is an assistant professor in the management department at Worcester Polytechnic Institute, Worcester, Mass.

Yang W. Lee (ylee@crgi.com) is associate director for the Total Data Quality Management Research Program at MIT, Cambridge, Mass., and president and CEO of Cambridge Research Group.

Richard Y. Wang (rwang@mit.edu) is co-director for the Total Data Quality Management Research program and associate professor at MIT Sloan School of Management, Cambridge, Mass.