Giri Kumar Tayi and Donald P. Ballou, Guest Editors

Examining Data Quality

Ensuring the quality of the data resource has been a continuing concern to those in the information systems profession. Over time techniques and procedures have evolved designed to make sure the data required by traditional transaction processing systems possesses an appropriate level of quality. However, the use of legacy data in, for example, decision and executive support systems has refocused attention on data quality and has exposed problems such as the need for "soft" data not encountered in traditional systems. Furthermore, data is now viewed as a key organizational resource and should be managed accordingly.

The term "data quality" can best be defined as "fitness for use," which implies the concept of data quality is relative. Thus data with quality considered appropriate for one use may not possess sufficient quality for another use. The trend toward multiple uses of data, exemplified by the popularity of data warehouses, has highlighted the need to address data quality concerns.

Furthermore, fitness for use implies that one needs to look beyond traditional concerns with the accuracy of the data. Data found in accounting-type systems may be accurate but unfit for use if that data is not sufficiently timely. Also, personnel databases situated in different divisions of a company may be correct but unfit for use if the desire is to combine the two and they have incompatible formats.

A related problem with multiple users of data is that of semantics. The data gatherer and initial user may be fully aware of the nuances regarding the meaning of the various data items, but that will not be true for all of the other users. Thus, although the value may be correct, it can easily be misinterpreted. Also, the capability of judging the reasonableness of the data is lost when users have no responsibility for the data's integrity and when they are removed from the gatherers. Such problems are becoming increasingly critical as organizations implement data warehouses.

Another trend that explains the heightened aware-

PAUL WATSON

xamining
Data
Qualit
1010001
C127
R245
T.P4
R246
C124

ness of the importance of data quality is the increased usage of soft data in computer-based systems. By soft data we mean data with inherently unverifiable quality—an example would be data regarding competitors' intentions. Management, of course, has always used soft data, but data stored in computer systems has traditionally been limited to hard data, which inherently is verifiable. To be truly effective, systems that support managers and executives must be able to handle soft data.

In a very real sense data constitutes the raw material for the Information Age. Unlike physical raw material, however, data is not consumed and in fact can be reused repeatedly for various purposes. What we call data may well be meaningless unless placed in some context. The value of raw material to an organization is clear, at least from an accounting perspective. The value of data, in contrast, depends almost entirely on its uses, which may not even be fully known.

It has long been recognized that data is best described or analyzed via multiple attributes or dimensions. Ballou and Pazer [1] identified and discussed four dimensions of data quality: accuracy, completeness, consistency, and timeliness. An example of the role of these dimensions can be found in Laudon's study of data problems in the criminal justice system [2]. Accuracy could refer to recording correctly facts regarding the disposition of a criminal case, completeness to having all relevant information recorded, consistency to a uniform format for recording the relevant information, and timeliness to recording the information shortly after the disposition. Imagine the impact on individuals of poor quality on any of these dimensions.

Wang and Strong recently analyzed the various attributes of data quality from the perspective of those who use the data [4]. Their analysis began by soliciting information from users regarding various quality descriptors attributable to data that resulted in over 100 items that were grouped into about 20 categories. These were further grouped into four broad data quality classes: intrinsic, contextual, representational, and accessibility. Accuracy belongs to intrinsic, completeness, and timeliness to contextual and consistency to representational data quality classes. Problems with data quality cannot be addressed effectively without an understanding of the data quality dimensions.

Difficulties Ensuring Data Quality

One can argue that ensuring the quality of data is much more difficult than is the case with manufactured goods. The raw material—the data—may well be of uncertain quality and its uses may be only partially known. The effectiveness of possible quality control procedures is uncertain if the data undergoes a series of ad hoc processing steps. It is possible technically to combine collections of data that were never meant to be combined. In addition to these difficulties there are other factors that complicate the job of a Data Quality Manager (DQM), an individual or group responsible for ensuring data quality.

The first factor concerns uncertainty as to what constitutes the data resource. The answer is neither obvious nor simple. One might be tempted to reply "all data stored in computers." What about engineering drawings? And word-processing documents? To narrow the scope, one might suggest certain key systems. But what if these systems use data from unreliable, external sources? Should the DQM attempt to work with those sources to improve the quality?

Another problem is the low priority often assigned to data quality. In some ways data quality assurance and computer security are analogous. Almost everyone agrees that ensuring computer security is an important activity, but at budget time it tends to get shortchanged. It has been said that nothing increases the budget for computer security like a well-publicized breach or disaster. Similarly, ensuring data quality is widely recognized as a valid and important activity, but in practice few people list it as a top priority. One of the major responsibilities of the Chief Information Officer and the DQM should be to sensitize executives and managers to the importance of ensuring data quality.

The problem of ensuring data quality is exacerbated by the multiplicity of potential problems with data. There are so many ways for data to be wrong. In a sense the foreign exchange rates in today's newspaper are timely, and yet in actuality the values were out of date before the newspaper was printed. For a particular file every data value could be both accurate and timely, but certain records (rows) may be missing entirely. The sales data may be accurate, timely and so forth, and yet be inconsistent and hence of little use if an inappropriate reporting period is used. No one can anticipate all the circumstances that could compromise the integrity of an organization's data. Awareness of this does not provide a solution but is a necessary ingredient for effective data quality management.

It is sometimes difficult to determine how serious are deficiencies with the data. For example, most people would agree that using different alphanumeric codes in different divisions of the organization to represent the same item is not desirable. Yet this need not create any difficulties. If these data items remain confined to the division of origin and are never combined across divisions, then there is no need to change anything. However, if the data items are made available across divisions on an ad hoc basis, as would be the case with a data warehouse, then something needs to be done to resolve the inconsistencies.

A somewhat similar problem is determination of the nature of data deficiencies. This is especially true in multiuser environments, for users may well have differing data quality requirements. It is necessary for the DQM to have an awareness of what could lead to inadequate data quality. A first step in understanding how data can go bad is to recognize the fact that data have multiple attributes or dimensions, as discussed earlier. A set of data may be completely satisfactory on most of these dimensions but inadequate on a critical few.

It is important but often difficult to determine the appropriate level of data quality. Although one might wish that all of the organization's data were perfect in every way, achieving that could bankrupt the organization. Nor is having such pure quality necessary. Fitness for use implies that the appropriate level of data quality is dependent on the context. Determining the needed quality is difficult when differing users have differing needs. One might be tempted to state that the use requiring the highest quality should determine the overall level of quality. But what if that use is rather minor and unimportant to the organization, whereas the major use of the data does not require anywhere near such a level of quality? Thus it is necessary to balance conflicting requirements for data quality.

Within the last several years both practitioners and academicians have been developing procedures, techniques and models for addressing problems involved in ensuring data quality. The articles in this special section highlight and extend several of the themes discussed here. It is of fundamental importance to have an overall plan or blueprint for ensuring the quality of information products. Such an overview is provided in "A Product Perspective on Total Data Quality Management," by Richard Wang, a prominent researcher in and articulate spokesperson for the field of data quality. The article combines various research endeavors with the experiences of practitioners to produce a framework for dealing systematically with data and information quality issues. The methodology presents a cycle parallelling that found in manufacturing and contains concepts and procedures that facilitate defining, measuring, analyzing, and improving data quality.

Ken Orr, a pioneer in the development of tools and techniques for information systems analysis and design, brings to bear his experiences and expertise in "Data Quality and Systems Theory." He presents concepts that are critical for ensuring an information system will generate outputs over time that are trusted and hence used. Further, he identifies six data quality rules and explores the implications of these rules. A recurring theme in his work is the need for continual feedback from users to ensure that the data's quality is maintained.

The article "Assessing Data Quality in Accounting Information Systems," by Kaplan, Krishnan, Padman, and Peters exemplifies the role of analytical modeling in information systems in general and data quality in particular. Such approaches facilitate the systematic exploration of data quality issues, and they are invaluable for obtaining relevant and valuable insights. They present a decision support system to assist auditors to carry out data quality assessments of accounting information systems. The focus of the proposed system is to enable the auditors to decide the extent of testing and to select the minimum set of control procedures needed to ensure data reliability.

Thomas Redman is a well-known practitioner in the field and the author of the first practice-oriented book devoted to data quality (see [3]). His article, "The Impact of Poor Data Quality on the Typical Enterprise" is aimed at sensitizing senior management to the consequences of poor data quality. Specifically, he details how poor data quality affects operational, tactical, and strategic decisions.

We hope the articles in this section help increase awareness of the significance of data quality and ultimately improve the quality of the data that is the foundation of your organization.

References

1. Ballou, D.P. and Pazer, H.L. Modeling data and process quality in multi-input, multi-output information systems. *Management Science 31*, 2 (1985), 150–162.
2. Laudon, K.C. Data quality and due process in large interorganizational record systems. *Commun. ACM 29*, 1 (1986), 4–11.
3. Redman, T.C. *Data Quality for the Information Age.* Artech House, Boston, MA, 1996.
4. Wang, R.Y. and Strong, D. Beyond accuracy: What data quality means to data consumers. *J. Manage. Info. Syst. 12*, 4 (1996), 5–34.

Giri Kumar Tayi (gk952@cnsibm.albany.edu) is an associate professor in the Department of Management Science and Information Systems at the State University of New York at Albany.
Donald P. Ballou (dpb67@cnsibm.albany.edu) is an associate professor in the Department of Management Science and Information Systems at the State University of New York at Albany.