The Information Quality of Databases

InduShobha Chengalur-Smith  
*University at Albany, USA*

M. Pamela Neely  
*Rochester Institute of Technology, USA*

Thomas Tribunella  
*Rochester Institute of Technology, USA*

**INTRODUCTION**

A database is only as good as the data in it. Transaction-processing systems and decision-support systems provide data for strategic planning and operations. Thus, it is important to not only recognize the quality of information in databases, but also to deal with it. Research in information quality has addressed the issues of defining, measuring and improving quality in databases; commercial tools have been created to automate the process of cleaning and correcting data; and new laws have been created that deal with the fallout of poor information quality. These issues will be discussed in the next sections.

**BACKGROUND**

Data or information quality began to achieve prominence during the total quality management movement in the 1980s. However, it has relevance to individual decision making as well as to organizational data. With the advent of the World Wide Web, the concept of information overload became universal, and the notion of information quality had instant appeal. As is typical in an emerging field, there are several research themes in the information quality literature. In this section, we highlight topics in some of the major research areas.

Although data quality is now widely perceived as critical, there is less agreement on exactly what is meant by high-quality information. Clearly it is a multidimensional construct whose measure is very much context dependent. A substantial amount of work has been devoted to identifying the dimensions of data quality and their interrelationships. Early work on the dimensions of data quality looked at pairs of attributes, such as accuracy and timeliness, and the trade-offs between them (e.g., Ballou & Pazer, 1995).

A more comprehensive study that attempted to capture all the dimensions of information quality was conducted by Wang and Strong (1996). They conducted a two-stage survey and identified about 180 attributes of data quality, which they combined into 15 distinct dimensions. These dimensions were further organized into four categories: intrinsic quality, contextual quality, representation, and accessibility.

There have been several follow-on studies that focus on subsets of these dimensions. For instance, Lee and Strong (2003) examined five of these dimensions and concluded that when data collectors know why the data is being collected, it leads to better quality data. Another avenue for research has been in developing formulae and metrics for each of these dimensions (e.g., Pipino, Lee, & Wang, 2002).

Yet another branch of research in information quality considers information to be a product of an information manufacturing process (Ballou, Wang, Pazer, & Tayi, 1998). Thus, the output of an information system can be tracked much like a manufacturing product, allowing for the application of quality-control procedures to improve the quality of information products. This concept has been elaborated on by Shankaranarayan, Ziad, and Wang (2003) by creating an information product map that represents the flow and sequence of data elements that form the information product.

**MAIN FOCUS**

Wang, Lee, Pipino, and Strong (1998) listed the steps an organization must take to successfully manage information as a product, which has led to the development of several software tools and methods for modeling and measuring data quality.

Not long ago, the majority of the software tools that fell into the data quality area depended on comparing data in databases to something else (Neely, 1998). Broadly defined, data quality tools provide three functions. They audit the data at the source; clean and transform the data in the staging area; and monitor the extraction, transformation, and loading process. Some of the
tools are an extension of data validation rules that should be a part of the database design but are frequently ignored in legacy systems. These tools, commonly referred to as auditing tools, involve comparing the data to a known set of parameters, which might be a minimum value, a set of dates, or a list of values. Once the data is run through the tool, results are routinely examined manually, a very time- and labor-consuming process.

Another category of tools, data cleansing tools, originally began as name and address tools. The core functionality of these tools is the ability to parse, standardize, correct, verify, match, and transform data. Ultimately, these tools are designed to produce accurate lists that can be used or sold.

Acquisitions and mergers have consolidated what was once a fragmented set of software tools that addressed specific roles in the data-quality process into integrated tools that can be used throughout the process of arriving at high-quality data. Companies such as Trillium Software, which once provided tools focusing on the audit function, have now expanded their functionality into the cleansing arena. Ascentials Software acquired Vality, primarily a data cleansing tool, and Ardent, primarily an auditing tool, providing it with the ability to monitor the entire process. Some of the earliest techniques in automating the process involved data mining techniques, discovering patterns in the data and reverse-engineering the process by suggesting business rules. WizRule has taken an early lead in using this technology.

Data profiling is being heralded as the new generation of data quality tools. Tools that provide this functionality, such as Data Quality Suite, claim that the inherent problem with auditing data is that it prevents the user from seeing the big picture. The focus is on the data elements, whereas with data profiling, the focus is on what the data should look like, based on aggregates of the data.

The metadata associated with data expands as the data is used for secondary purposes, such as for data warehousing, which in turn supports analytical tools, such as data mining or On-Line Analytical Processing (OLAP) tools. What was once a data dictionary, describing the physical characteristics of the data, such as field type, becomes a complex web of physical definitions, quality attributes, business rules, and organizational data. Some of the newest tools are those that attempt to manage metadata. Commercial metadata tools fall into two categories: tools for working with metadata, and centralized repositories. The metadata tools attempt to interpret technical metadata so that the business users can understand it (Levin, 1998). Central repositories are based on industry standards, in an effort to allow third-party warehouse and analysis tools to interface with them. Many systems attempt to combine the technical specifications with business information (Platinum Technology, 2000); however, there is still room for improvement.

Information quality has also been studied as a contributing factor to information systems (IS) success. A widely accepted model of IS success holds that the quality of information obtained from a system impacts satisfaction with, and ultimately the success of, the system (DeLone & McLean, 2003). In an empirical study of the factors affecting the success of data warehouse projects, data quality is shown to have a significant impact (Wixom & Watson, 2001). However, organizational, project and technical success did not result in high-quality data in the warehouse. A study by Strong, Lee, and Wang (1997) identified the major reasons for poor-quality data and suggested ways to improve the data collection and information dissemination processes.

The major impact of poor-quality data is on decision making. To make users aware of the quality of the data they are dealing with, the use of data tags has been proposed (Wang & Madnick, 1990). Databases could be tagged at varying levels of granularity and using different dimensions of data quality. For instance, by recording the time a data item was last updated, a user could assess the currency of the data. Creating separate tags for each data item would be very resource intensive. Alternatively, a tag for each record or field could be used to represent the quality of the information stored there. When considering what information the tag should contain, it is important to remember that the data will be used in the context of a specific situation. Thus, the quality of the data may be sufficient in one instance, but not in another. The “fitness-for-use” concept continues to be an issue when applying data tags. Regardless, the data tags form a metadata about the quality of the data.

Clearly, keeping the metadata on information quality up to date could itself be a Herculean task. Before embarking on such an ambitious project, it would be worthwhile to examine the impact of providing this information about the quality of the data. If decision makers were unable to use this information, for any reason, whether it was because it was not in the appropriate format or because it created information overload, then the effort of producing this metadata would be futile.

Some exploratory research (Chengalur-Smith, Ballou, & Pazer, 1999) has shown that the format in which the data quality information is provided does play a role in the use of the information. Fisher, Chengalur-Smith, & Ballou, (2003) determined that experienced decision makers were more likely to use the data quality information if they did not have expertise in the data domain and if they did not feel time pressure. Further research will refine these findings.