

A Multidisciplinary View of Data Quality

Andrew Borchers

Kettering University, USA

M

INTRODUCTION

This article introduces the concepts of data quality as described in the literature of several disciplines and discusses research results on how individual perceptions of data quality are influenced by different media (in particular World Wide Web vs. print). A search of literature on “data quality” and “media creditability” reveals that researchers in many disciplines are separately studying the subject. These disciplines include accounting, advertising and public relations, information systems, scientific data collection, education, journalism fields, and others. While these threads have developed separately, these streams of research approach similar issues of how people view the quality of information they receive from different sources.

BACKGROUND

Data quality is an emerging area of research fundamental to the field of information systems. Indeed, the efficacy of systems is in large part driven by the quality of the data that they contain. With the Internet revolution, however, there have been fundamental changes in how information is collected and shared that have a potentially great influence on data quality. This challenge is accentuated with the recent move to “user-generated content” as a part of the broader evolution to Web 2.0 (Schwartz, 2007). In addition, younger generations immerse themselves in media more than their parents do. This has led to the label of the “M-generation.” A study by the Kaiser Family Foundation and Stanford University finds young people spending on average 6.5 hours per day in media exposure. Increasingly, this exposure comes in multiple media at one time (Azzam, 2006).

However, with such access and participation comes a challenge as stated by Gilster (as cited in Flanigan & Metzger, 2000):

When is a globe spanning information network dangerous? When people make too many assumptions about what they find on it. For while the Internet offers myriad opportunities for learning, an unconsidered view of its contents can be misleading and deceptive.

Further, organizational responses to data quality have been largely ad hoc (Swartz, 2006) with the majority of

firms relying on localized, ad hoc approaches to ensuring data quality.

Recent research and seminars underscore the importance of the topic of data quality. Interest in the discipline has spawned the creation of the International Association for Information and Data Quality, several annual conferences (e.g., www.iqconference.org), and the *ACM Journal of Data and Information Quality*. Indeed, Total Data Quality Management (TDQM) has evolved as a field of study extending the concepts of Total Quality Management (Radziwill, 2006). Data quality has emerged as a significant research area.

Information systems and journalism practitioners have echoed the importance of data quality for many years. Research by Redman (1998) summarizes the practical implications of poor data quality. He points out the consequences of poor data quality in areas such as decision making, organizational trust, strategic planning and implementation, and customer satisfaction. Redman conducted (1998) detailed studies and found increased costs of 8-12% due to poor data quality. Service organizations can find increased expenses of 40-60% (Redman, 1998). Strong, Lee, and Wang (1997) support the seriousness of this issue in their study of 42 data quality projects in three organizations. Early research by other authors note data quality issues in a number of settings including accounting (Xu, 2000; Kaplan, Krishnan, Padman, & Peters, 1998), airlines, healthcare (Strong et al., 1997), criminal justice (Laudon, 1986), and data warehousing (Ballou, 1999).

As for a formal definition of data quality, Umar, Karabatis, Ness, Horowitz, and Elmagardmid (1999) quote Redman (1992):

A product, service, or datum X is of higher quality than product, service, or datum Y if X meets customer needs better than Y.

Umar et al. (1999) go on to point out that this definition has been generally accepted and is consistent with the author’s work. The definition is somewhat incomplete, however, as it does not delve into the various dimensions of data quality.

A number of authors in the information systems field have gone further than Redman and written conceptual articles on “data quality” (Wand & Wang, 1996; Wang, Reddy, & Kon, 1995; Wang & Strong, 1996; Strong et al., 1997). This work suggests that data quality is a multidimensional

concept (Wand & Wang, 1996) that researchers can view from a number of different perspectives. A panel discussion in 2000 (Lee, Bowen, Funk, Jarke, Madnick and Wand) found five different perspectives to discuss data quality. These included an ontological perspective (specification of a conceptualization) that included different views of reality based on actual observation vs. computer-influenced observations; an architectural perspective, a view that focuses on system infrastructure and its influence on data quality; a context mediation perspective, focusing on communication across space and time; a time-based e-commerce perspective, focusing on the real-time nature of e-commerce; and an information product perspective, focused on data as a product of an organization.

In talking about “data quality,” a key beginning is to determine from the literature just what one means by the term. In a definitive work on the topic, Wang and Strong (1996) provide a conceptual framework for data quality. In a way consistent with Redman’s (1992) customer perspective, they start by defining “high-quality data as data that is fit for use by data consumers.” Using a two-stage survey and sorting process, Wang and Strong (1996) develop a hierarchical framework for data quality that includes four major areas: intrinsic, contextual, representational, and accessibility.

Intrinsic data quality refers to the concept that “data have quality in their own right” (Wang & Strong, 1996). Intrinsic dimensions include accuracy, objectivity, believability, and reputation. Contextual data quality is based on the idea that data does not exist in a vacuum—it is driven by context. Contextual dimensions include relevancy, timeliness, and appropriate amount of data. Representational data quality relates to the “format of the data (concise and consistent representation) and meaning of data (interpretability and ease of understanding).” Accessibility data quality refers to the ease with which one can get to data (Wang & Strong, 1996).

More recent research reinforces many of the concepts presented above. In information systems research, data quality is of particular interest to work on data warehouses and business intelligence. In a recent article noting “BI at age 17” (Martens, 2006), Howard Dresner, author of the term “business intelligence,” notes the importance of data quality due to its impact on business process management and operational planning. In studying the maturity of data warehouse projects, Sen, Sinha, and Ramarmuthy (2006) note that data quality is a key determinant. Crie and Micheaus (2006) note that data quality management is a key step in the customer data to value information chain.

Beyond the information systems literature, journalism provides a second relevant body of literature. One of the focus points is on perceptions of Internet credibility (Flanigan & Metzger, 2000; Johnson & Kaye, 1998; Bucy, 2003). The major thrust of this literature is in comparing the Internet to traditional sources with respect to credibility. Note that

when referring to “credibility,” these authors say “the most consistent dimension of media credibility is believability, but accuracy, trustworthiness, bias and completeness of information are other dimensions commonly used by researchers” (Flanigan & Metzger, 2000, p. 521). Hence, there is a rough correspondence of thinking about “credibility” in the journalism literature to the concept of “intrinsic” and “contextual” data quality in the information systems literature. One author in this field (Bucy, 2003) goes on to differentiate “media” credibility from “source” credibility and suggests that researchers have viewed these two forms of credibility as being separate areas of research.

A third field has contributed to the same discussion, namely, advertising and public relations research. Working on a variety of topics, researchers have asked the question: “What impact does media credibility have in [an] organization’s advertising and public relations efforts?” Huh, DeLorme, and Reid (2004) studied media credibility in the context of direct-to-consumer prescription drug advertising. They examined consumer perceptions of credibility based on age and media. Greary (2005) studied the impact on the public relations field of declining media credibility, reported in 2004 to be at a 30-year low. Finally, Cable and Yu (2006) studied job seekers and their organizational image beliefs of potential employers. In their work, they considered three different recruitment media and found media richness to be associated with job seekers’ image beliefs.

The concept of data quality also appears in disciplines such as accounting and finance. With the passing of Sarbanes-Oxley legislation and an increased focus on the accuracy financial reporting, new standards focused on data quality are emerging. Clark (2006) reports on the adoption of corporate action standards and points out that data quality is a significant concern. Schwarzkopf (2007) examines source credibility and investors’ attitudes toward financial and non-financial performance measures. Interestingly, he noted no difference between more and less experienced investors. He did note, however, that source credibility was most important to investors when viewing financial estimates compared to non-financial performance measures.

In yet another discipline, that of scientific data collection, similar dimensions appear. Radziwill (2006) quotes Loshin (2001) in dividing data quality into four areas: data models, data values, information domains, and data presentation. Within each of these four areas, Loshin gives further dimensions that are quite similar to Wang and Strong’s (1996) work. In a similar fashion, Radziwill (2006) also quotes Graefe (2003) in describing data quality criteria in the context of decision process.

It is interesting to note how authors working in multiple disciplines have chosen many of the same dimensions in speaking about data quality. Table 1 summarizes the dimensions these authors have identified using Wang and Strong’s (1996) framework.

5 more pages are available in the full version of this document, which may be purchased using the "Add to Cart" button on the publisher's webpage: www.igi-global.com/chapter/multidisciplinary-view-data-quality/13975

Related Content

New Forms of Collaboration & Information Sharing in Grocery Retailing: The PCSO Pilot at Veropoulos

Katerina Pramati and Georgios I. Doukidis (2005). *Journal of Cases on Information Technology* (pp. 63-78). www.irma-international.org/article/new-forms-collaboration-information-sharing/3162

Breaking Out of Lock-In: Insights from Case Studies into Ways Up the Value Ladder for Indian Software SMEs

Abhishek Nirjar and Andrew Tylecote (2005). *Information Resources Management Journal* (pp. 40-61). www.irma-international.org/article/breaking-out-lock/1280

CAL Student Coaching Environment and Virtual Reality in Mechanical Engineering

S. Manjit Sidhu, N. Selvanathan and S. Ramesh (2008). *Information Communication Technologies: Concepts, Methodologies, Tools, and Applications* (pp. 1696-1711). www.irma-international.org/chapter/cal-student-coaching-environment-virtual/22770

Embracing Change: How South Ayrshire Council Library Service Became a World Leader in Electronic Resources Management

Jean Inness (2014). *Progressive Trends in Electronic Resource Management in Libraries* (pp. 210-220). www.irma-international.org/chapter/embracing-change/90183

Modeling Back Office Operations at Greenfield Online's Digital Consumer Store

Gerald C. Campbell, Christopher L. Huntley and Michael R. Anderson (2003). *Annals of Cases on Information Technology: Volume 5* (pp. 358-369). www.irma-international.org/chapter/modeling-back-office-operations-greenfield/44552