# Visible Evidence of Invisible Quality Dimensions and the Role of Data Management

Ayoung Yoon[1]
[1]Indiana University Purdue University, Indianapolis

**Abstract**
Past research has shown that data reusers are concerned with the issue of data quality and the identified attributes of quality. While data reusers find evidence of the attributes of data quality during their assessment of data for reuse, there may be other dimensions of data quality that reusers are concerned about but that are not always visible to them. This study explores these invisible dimensions of data quality that have been identified by data reusers. The findings of this study indicate that data reusers are concerned with two kinds of invisible characteristics for assessing the data: the efforts put on data, and the ethics behind the data. While these quality dimensions cannot be easily measured at face-level, data reusers find proxy evidence that indicates the presence of these invisibilities. This finding signifies the role of data management that can make these invisible data qualities visible.
**Keywords:** data reuse; data curation; trust
**doi**: 10.9776/16123

**Contact**: ayyoon@iupui.edu

## 1 Introduction

With considerable recent attention given to the data deluge[1] and scientific and interdisciplinary research, researchers anticipate the future reuse of valuable data. The benefits of data reuse, which have been well discussed by researchers, include the following: data reuse enables others to ask new questions of extant data, advance solutions for complex human problems and the state of science, reproduce research, and expand the instruments and products of research to new communities (Borgman, 2010; Borgman, 2011; Hey & Trefethen, 2003; Hey, Tansley, & Tolle, 2009). As such, researchers' interest in reusing existing data is growing.

The reuse of existing data requires researchers to carefully assess the data for its quality. Past research has shown that reusers are concerned with this issue (e.g., Cragin & Shankar, 2006; Zimmerman, 2008). Various attributes of data that reusers consider important were identified, including relevancy, data validity, documentation quality, and reputation of data producers (Faniel & Jacobsen, 2010; Van House, 2002; Van House et al., 1998), which can be dimensions of data quality from users' perspectives. Data reusers try to find evidence of these attributes to confirm that the data meets their quality criteria. While data reusers may find visible evidence of these attributes, there could be other dimensions that data reusers are concerned with but that are not always visible to them. In this paper, I examine the invisible dimensions of data quality for which data reusers try to search during their examinations. The research question motivating this study is as follows: What are the invisible dimensions of data that data reusers are concerned with in terms of quality assessment? How do data reusers try to find evidence of these invisible dimensions?

The study data includes 38 interviews with quantitative data reusers within the social science domain (public health and social work). The findings of this study indicate that data reusers are concerned with two invisible characteristics of data for assessing data: the efforts put on data, and the ethics behind the data. While these quality dimensions cannot be easily measured at face-level, data reusers find proxy evidence that indicate these invisibilities. This finding signifies the role of data management that can make these invisible qualities of data visible.

## 2 Literature Review

### 2.1 Data Reusers' Criteria for Quality Data

The topic of data quality has received a great deal of recent attention in official documents such as the Data Quality Act (Public Law 106-554; H.R. 5658, Sec. 515), the Office of Management and Budget (OMB)'s guidelines (2002), and Information Quality Guidelines by National Science Foundation (NSF).

[1] The term was first introduced by Hey and Trefethen in 2003.

Recent discussions have expanded on the definition of data quality and acquisition, curation and assuring of high-quality data (e.g., Hense & Quadt, 2011; RIN, 2008; Marchionini et al., 2012). Although common definition of quality includes the terms "fitness for use" (e.g., Madnick, et. al 2009; Wang & Strong, 1996), "value of information," "reliability," and "validity" (e.g., Altman, 2012; Raykov & Marcoulides, 2010), there has yet to be agreement on the definition. Ashley (2012) argues that it may be possible to agree only that quality is desirable. Discussions on quality attributes have been recently initiated using the influence of information quality research (e.g., Arazy & Kopak, 2011; Knight & Burn, 2005; Lee at al., 2002) and MIS research (e.g., Madnick et al., 2009; Fox et al., 1994; Wang & Strong, 2006). Examples of the adopted or identified attributes of quality, as taken from recent discussions, include integrity, usability, objectivity, understandability, authenticity, accuracy, comprehensiveness, utility, transparency, and accessibility (Giarlo, 2012; OMB, 2002; Sticco, 2012).

The most important consideration defining the quality of data is that it is defined by users' needs, as users can request quality as an end-consumer of the data. Acknowledging this importance, researchers have integrated the users' perspective. The OMB's guidelines (2002) offer a use-related dimensional definition, explaining quality as an encompassing term comprising integrity, objectivity, utility, and usefulness to its intended users. Marchionini et al. (2012) reported that users define data quality based on the data properties required for use in scientific research. More research is desired to broaden the understanding of quality dimensions, particularly from users' perspectives.

### 2.2 Data Management and Quality

Data quality has been a concern of data management research because a loss of data quality during the management process reduces the ways in which the data can be adequately used (Martin & Ballard, 2010). For instance, the generation and extraction of contextual data information (metadata), which are key parts of data management (Giarlo, 2012), are considered to be one quality dimension (Marchionini et al., 2012); actions such as checksum, replication, media refreshment, version management, and prevention of unauthorized access and corruption can ensure that data retains its integrity and is not altered or destroyed in unsanctioned ways (Giarlo, 2012). Fixity computation, auditing, and the detection of storage corruption are equally important because it affects data quality properties (Altman, 2012). The long-term preservation of data is not only related to issues of authenticity, chain of custody, and integrity; it is often associated with the expected value of future use and increased "fitness for use" (Altman, 2012). These examples show that data management activities interact with the data quality in distinctive ways and the role of data management is significant in ensuring quality attributes: data management supports quality through the selection of data that fits the needs of designated communities and the descriptions or assertions that demonstrate an appropriate management of uncertainties surrounding the data through integrity and provenance checks. While the meaning of "quality" in a management context can differ from the users' meaning, based on the relationship between data management and quality presented from the past research, it is important to understand how data management can support quality from users' perspectives.

## 3 Methods

### 3.1 Study Sample and Recruitment

This study addressed one specific type of data reusers: quantitative social science data reusers from the public health and social work domains. The study participants were identified from the major scholarly databases – including EBSCOHost, SAGE Journals, ProQuest Social Science, and ERIC – by searching for "secondary data" and "secondary analysis." The searches were limited to journals and conference proceedings published in English. I chose public health and social work for several reasons: both disciplines have listed enough data reusers for this study; both disciplines have a professional orientation in their research; and some data sets were used by both disciplinary researchers. All of these helped me to recruit a homogeneous sample.

### 3.2 Data Collection and Analysis

Among the 229 initially identified potential participants, 38 data reusers affirmatively responded and were interviewed. A semi-structure interview protocol solicited interviewees' perceptions and thoughts from their data reuse experiences. Given the diverse geographic locations of the participants, phone interviews were conducted. The average length of the interviews was 60 minutes, and each interview was recorded and fully transcribed. The data was analyzed using the qualitative data analysis software Nvivo 10 for Mac through iterative and inductive cycles.

# 4 Findings

## 4.1 Research Participants

The interviewees were all researchers in various positions (PhD student, post-doc, assistants to full professors, and research scientists), with a mix of genders (male: 13; female: 25) and ages (ranging from their 20s to 70s). Several participants had obtained and reused research data from institutions, including data from federal and state government organizations, which are mostly publicly available, and research data from individuals or individual research teams. Eleven participants had only reused data from institutions; seven participants only reused data from individual researchers or research teams; and the remaining 20 used both types of data in their research. Data repositories were engaged in the process of acquiring data from institutions and individual researchers for four of the participants.

## 4.2 General Dimensions of Quality

The participants mentioned several attributes of quality data during the interviews. While it is not the focus of this study, it is important to understand the general quality dimensions discussed by participants, as it will help contrast the invisible dimensions of quality data. Participants usually found direct evidence of the general quality dimensions – e.g., by the topic, number of sample, formats of data, study design, and documentation.

- Relevancy: Data should meet the research needs (e.g., "It had a best measure of what I'm looking for [IP13]") or be relevant to the topic of research (e.g., "all the variables in the study answered my questions [IP02]").
- Representative or large sample: This may be a distinctive attribute of quantitative data, particularly in domains that feature an emphasis on large samples. Data should have a representative national sample to seek generalized implications for a stronger claim, use nations as the unit of analysis, or compare the national estimates to the samples.
- Validity: Data validity is an object quality of data and a core part. Good methodology, good measurement, and a valid scale for further reliability were considered to be part of validity and a quality dimension.
- Usability: Good data should be easily accessible and usable. Data formats and proprietary software influenced the usability of data.
- Understandability: Good quality data should be understandable and supported by good quality documentation, easy access to the original study's principle investigators, and necessary help provided to use the data.

## 4.3 Invisible Dimensions of Quality and Visible Evidence

While participants usually found evidence of the general quality dimensions, they pointed out two additional dimensions of quality data that are less visible at face-level: efforts put on data, and the ethics behind the data. These invisible dimensions are related to the characteristics of the humans behind the data and are usually perceived as data producers by participants.

### 4.3.1 Efforts and Commitment

The first dimension of data that participants cared about was data producers' efforts and commitment to the data. Participants considered data producers' efforts and commitment as a way to improve and/or guarantee the quality of data. On a general level, the participants believed the relationship between these efforts and the quality results coming from these efforts: IS17 noted, "[W]hen you spend times and put your efforts into something, it's less likely [to go] wrong." In addition, the efforts and commitment of the researchers in their data reflected their attitude on the research and data of the participants. IS06 mentioned the impact of data producers' attitude on research as well as data from the research and criticized data from "any researcher [does] like, 'Just get it done,' not paying enough attention in study design, measurement, things like that." IS06 believed, "rigorous research cannot be done without cares and deep thinking. (. . .) Same as data, the level of quality [of] data is not achievable without same amount of cares." Despite participants' belief that the efforts of data producers contribute to quality data, it was not directly visible for the participants because effort is a human attempt, not a data attempt. Participants, thus, tried to find alternative evidence as proxies of the human efforts put on the data.

#### Evidence 1. Preparedness

Participants considered the level of preparedness of the data, which can be different depending on the participants' views and expectations. In general, the participants discussed the preparedness of the data

at the preservation level in entire data packages; for instance, how data appeared on first impression, how data files are organized, and how data file names are assigned. The participants agreed that well-prepared data required a lot of effort from those who managed and prepared the data (e.g., "Everything was sort of in its place. It was clear that a lot of work and a lot of time had been spent putting all of [the files] together" [SI08]).

Evidence 2. Documentation

Documentation was further evidence of effort and is key to understanding and using data because it usually includes necessary information about the data, such as study design, data collection, principle investigators, and funding information. Participants gave more weight on the documentation to judge the efforts put on the data. Documentation that is "easy to follow and straightforward" (IS10) clearly showed that the data producers put out "hard work on data" (IP01). The participants often connected good documentation with extensive effort made on the quality of data: "You can tell from the documentation whether or not a research[er] was thorough and careful" (IS08).

Evidence 3. Fewer Errors and Mistakes

The final evidence was the inclusion of fewer errors and mistakes in the data. Participants generally understood that some human errors would be present in the data collection, cleaning process, and preparation, as "it was impossible to expect all the data (. . .) to be perfect" (IP18). However, participants interpreted reoccurring or frequent errors and mistakes as carelessness on the part of the data producer.

#### 4.3.2 Ethics and Intentions

The second invisible dimension of data includes the ethics and intentions behind the data. Ethics beyond the data creation was an important consideration for the participants, as they believed good science should comply with research ethics. IU18 believed, "The researchers all know things like honesty, objectivity, integrity . . . and have to conform with them in research, [and be] accountable to the public." Because researchers' ethics are inevitably reflected in the data, the participants expected the data producers to create data without a conflict of interest, collect and manage data ethically, and refrain from manipulating the data. The participants were also concerned with the intentions behind the data collection, whether the data was created for "science, scientific research, (. . .) public good" (IU17) or for the profit of an agency. While the participants could infer the intentions of the data's creation from the funding information, the ethical dimension of the researchers and the data was not always visible in the data itself.

Evidence 1. Openness and Transparency

Proxy evidence of ethics in data was openness and transparency of data. Participants expected nothing hidden in ethical research and about data: "Why not make all that information available if you do everything ethically and properly?" (IP04). Information about data should be open and transparent, including the original study description, data cleaning process, any changes made to the data, limitations, even issues with the data that data producers experienced. Especially when the original study was transparent about the limitation and issues within data (both technical and methodological), participants assumed that the data producers were ethical enough to report these: "They were very transparent about what they're doing," said IP15 which made IP15 "feel good" about data.

## 5 Discussion and conclusion

The findings of this study present the general and invisible quality dimensions identified by the study participants. While the general dimensions of quality show similarities with the previously identified quality attributes by past research, this study adds invisible dimensions, which are the human factors: who is involved in study design, the data collection, and the data management. Although data reusers are concerned with some data attributes such as validity, they also care about the human beings behind the data. This may not be true for other types of data, such as computational data where less human characteristics are involved; but within the social science context, the characteristics of the human beings behind the data are the hidden factors that influence the data quality. These characteristics, particularly the efforts and ethics of the researchers, were not always visible to data reusers from the data with which they interacted, and thus, the reusers searched for alternative evidence as a proxy: preparedness of data, documentation, and fewer errors as the evidence of efforts; and openness and transparency as the evidence of ethics.

It is not new that data management contributes to data quality, for instance, by enhancing the usability and understandability of the data, which is part of the general quality dimension identified by this

study's participants. Beyond that, data management contributes to the invisible dimensions of quality and helps to make these invisible dimensions more visible to data reusers. The proxy evidence of invisible dimensions that reusers find is all from the outcome of managed data. Data management helps with the proper naming of data files, file organization, the process of data documentation, and continuous quality control and assurance for preventing errors within data package (Martin & Ballard, 2010), which can be used as evidence of invisible quality dimensions identified by the study participants. Good documentation practices that meet the users' needs can increase the openness and transparency of data.

This study contributes to the understanding of data quality from users' perspectives and adds the ethical and effort dimensions to the quality. Even though data management aims to maintain high quality data through various management activities, it is important to know that data management also supports the invisible dimensions of quality providing evidence of human efforts and ethics.

## 6 References


Altman, M. (2012). Mitigating threats to data quality throughout the curation lifecycle. In *Curating for Quality: Ensuring Data Quality to Enable New Science* (Workshop Report) (pp. 20-31). Retrieved from http://datacuration.web.unc.edu/

Arazy, O., & Kopak, R. (2011). On the measurability of information quality. *Journal of the American Society for Information Science and Technology*, *62*(1), 89–99. doi:10.1002/asi.21447

Ashley, K. (2012). Generic data quality metrics-What and why. In *Curating for Quality: Ensuring Data Quality to Enable New Science* (Workshop Report) (pp. 89-92). Retrieved from http://datacuration.web.unc.edu/.

Borgman, C. L. (2010). Research data: Who will share what, with whom, when, and why? Presented at the *China-North America Library Conference*, Beijing. Retrieved from http://works.bepress.com/borgman/238

Borgman, C. L. (2011). The conundrum of sharing research data. *SSRN eLibrary*. Retrieved from http://papers.ssrn.com/sol3/papers.cfm?abstract_id=1869155

Cragin, M. H., & Shankar, K. (2006). Scientific data collections and distributed collective practice. *Computer Supported Cooperative Work (CSCW)*, *15*(2-3), 185–204. doi:10.1007/s10606-006-9018-z

Data Quality Act (Public Law 106-554; H.R. 5658, Sec. 515). Retrieved from http://www.ftc.gov/ogc/sec515/.

Faniel, I. M., & Jacobsen, T. E. (2010). Reusing scientific data: How earthquake engineering researchers assess the reusability of colleagues' data. *Computer Supported Cooperative Work (CSCW)*, *19*(3-4), 355–375. doi:10.1007/s10606-010-9117-8

Fox, C., Levitin, A., & Redman, T. (1994). The notion of data and its quality dimensions. *Information Processing & Management*, *30*(1), 9–19. doi:10.1016/0306-4573(94)90020-5

Giarlo, M. (2012). Academic libraries as data quality hubs. In *Curating for Quality: Ensuring Data Quality to Enable New Science* (Workshop Report) (pp. 20-31). Retrieved from http://datacuration.web.unc.edu/

Hense, A., & Quadt, F. (2011). Acquiring high quality research data. *D-Lib Magazine*, *17*(1/2). doi:10.1045/january2011-hense

Hey, T., & Trefethen, A. (2003). The data deluge: An e-Science perspective. In F. Berman, G. Fox, & T. Hey (Eds.), *Grid Computing* (pp. 809–824). John Wiley & Sons, Ltd. Retrieved from http://onlinelibrary.wiley.com/doi/10.1002/0470867167.ch36/summary.

Hey, T., Tansely, S. & Tolle, K. (Eds.). (2009). *The Fourth Paradigm: Data-Intensive Scientific Discovery*. Redmond, WA: Microsoft. Retrieved from http://research.microsoft.com/en-us/collaboration/fourthparadigm/4thparadigm_science.pdf.

Knight, S., & Burn, J. (2005). Developing a framework for assessing information quality on the World Wide Web. *Informing Science*, *8*, 159–172.

Lee, Y. W., Strong, D. M., Kahn, B. K., & Wang, R. Y. (2002). AIMQ: a methodology for information quality assessment. *Information & Management*, *40*(2), 133–146. doi:10.1016/S0378-7206(02)00043-5

Madnick, S. E., Wang, R. Y., Lee, Y. W., & Zhu, H. (2009). Overview and framework for data and information quality Research. *Journal of Data and Information Quality*, *1*(1), 2:1–2:22. doi:10.1145/1515693.1516680

Marchionini, G., Lee, C. A., Bowden, H., & Lesk, M. (2012). *Curating for Quality: Ensuring Data Quality to Enable New Science* (Workshop Report). Retrieved from http://datacuration.web.unc.edu/.

Martin, E., & Ballard, G. (2010). *Data Management Best Practices and Standards for Biodiversity Data Applicable to Bird Monitoring Data*. U.S. North American Bird Conservation Initiative Monitoring Subcommittee. Retrieved from http://www.nabci-us.org/aboutnabci/bestdatamanagementpractices.pdf

National Science Foundation. (n.d.). *Information Quality Guidelines*. Retrieved from http://www.nsf.gov/policies/nsfinfoqual.pdf.

Office of Management and Budget. (2002). Guidelines for ensuring and maximizing the quality, objectivity, utility, and integrity of information disseminated by federal agencies; Notice; Republication. *Federal Register*, *67*(36). Retrieved from http://www.whitehouse.gov/sites/default/files/omb/fedreg/reproducible2.pdf.

Raykov, T., & Marcoulides, G.A. (2010). *Introduction to Pscychometric Theory*, Routledge Academic.

Research Information Network. (2008). *Stewardship of digital research data: A framework of principles and guidelines*. Retrieved from http://www.rin.ac.uk/our-work/data-management-and-curation/stewardship-digital-research-data-principles-and-guidelines.

Sticco, J. C. (2012). Towards Data Quality Metrics Based on Functional Requirements for Scientific Records. In *Curating for Quality: Ensuring Data Quality to Enable New Science* (Workshop Report) (pp. 20-31). Retrieved from http://datacuration.web.unc.edu/

Van House, N. A., Butler, M. H., & Schiff, L. R. (1998). Cooperative Knowledge Work and Practices of Trust: Sharing Environmental Planning Data Sets. In *The ACM Conference On Computer Supported Cooperative Work* (pp. 335–343). Seattle, Washington.

Wang, R. W., & Strong, D. M. (1996). Beyond accuracy: What data quality means to data consumers. *Journal of Management Information Systems*, *12*(4), 5–33.

Zimmerman, A. S. (2008). New knowledge from old data: The role of standards in the sharing and reuse of ecological data. *Science, Technology & Human Values*, *33*(5), 631–652. doi:10.1177/0162243907306704