Selecting data quality dimensions: towards a business impacts assessment

Marco Spruit

Department of Information and Computing Sciences, Utrecht University, The Netherlands
m.r.spruit@uu.nl

As data volumes within enterprises grow, the number of errors in stored data and the organizational impact of these errors is likely to increase. CIOs and business executives must be able to justify the expense of each data quality initiative and convey the value proposition effectively to senior management. In order to do this, data quality needs to be expressed in terms of costs and organizational consequences, in order to be able to convey the value of improving data quality correctly. This research provides the first step towards this goal by determining the most frequently occurring data quality characteristics based on a comparative literature study of twelve frameworks.

1. Introducing the impacts of poor data quality in organizations

Data volumes within enterprises grow at a very high pace and enterprises are increasingly dependent on the timely availability of high quality data. In fact, many organizations' basis for competition has changed from tangible products to intangible information.

Poor quality information can have significant social and business impacts and there is strong evidence that data quality problems are becoming increasingly prevalent in practice (Wang and Wang, 1996). Most organizations have experienced negative effects of decisions based on information of inferior quality. Information quality issues have become important for organizations that want to perform well and obtain competitive advantage.

The DataWarehousing Institute (TDWI) estimates that poor quality customer data alone cost U.S. businesses over $600 billion a year. However, these data quality issues are often either not seen or ignored by most executives. According to TDWI's Data Quality Survey, almost half of all companies have no plan for managing data quality.

As the data volumes within enterprises grow, the number of errors in stored data and the organizational impact of these errors is likely to increase (Klein, 2002). More and more organizations do believe that quality information is critical to their success. However, not many of them have turned this belief into effective action. Enterprises seem reluctant to address, solve and prevent data quality issues until it is too late, making it a reactive process. The reason for this seems to be twofold: Management either accepts the status quo of their data environment as normal and acceptable, or they are unaware of the actual costs of poor quality data. Identifying the costs of poor data quality currently is indeed a cumbersome task. Creating a business case for fixing an organization's data environment has proven to be difficult.

Part of that difficulty is that data quality efforts are competing with other initiatives for IT budget dollars and staffing. CIOs and business executives must be able to justify the expense of the initiative and convey the value proposition effectively to senior management. In order to do this, data quality needs to be expressed terms of costs and organizational consequences to be able to convey the value of improving data quality correctly. There is currently no clear view on how data quality affects an organization as a whole, which makes expressing the added value of data quality improvement initiatives such a hard task.

Quantifying data quality improvement is a way to convince companies that steps should be taken to improve data quality throughout their business. In order to quantify data quality improvements, a thorough understanding of data quality itself is needed. This research further clarifies the term data quality by investigating its characteristics and its impact on organizations through the following research question: *"How can we determine the most relevant data quality characteristics with respect to assessing their business impacts in organizations?"*

The remainder of this paper is structured as follows. Section 2 surveys the identification and selection of data quality dimensions in existing frameworks. Section 3 interprets these findings in an ordered list of data quality dimensions based on the number of occurrences in the literature. Finally, section 4 concludes and discusses our further research direction.

2. Reviewing existing data quality frameworks

Data quality (DQ) can be considered as the quality of data values, or in other words the accuracy of those values. This has long been the view on data quality in practice. However, an investigation of data quality literature reveals many other characteristics of data quality (or information quality) than the mere accuracy of data values.

Definitions of quality found in literature and practice can, in general, be described as coming from either product-based or service-based perspectives. The product-based approach, commonly called data quality, focuses on the design and internal information systems view, and defines quality as the degree data satisfies initial specified requirements or the degree to which the data corresponds with real-world entities and facts. Typical criteria to measure the quality include completeness and accuracy of data. The issue with this approach is that there can still be deficiencies with respect to the initial specification of requirements of the data and the actual use of the data. This in turn has lead to a service-based approach to quality, commonly called information quality, which focuses on the information consumer and the consumer's use of the data. Using the term information instead of data implies that the delivery and use of data must be considered when one judges quality.

Information quality has been defined differently by several authors, but examining these definitions reveals a consensus about what information quality is. Kahn et al. (2002) define information quality as the characteristic of information to meet or exceed customer expectations, and as information that meets 'specifications' or 'requirements'. Other authors also describe in-formation quality as information that is most useful to the information customer. Fitness of use seems to be the most appropriate way to describe information quality and coincides with Juran's widely accepted definition of quality from 1974.

In order to evaluate information quality, many researchers have formulated key characteristics, often described as dimensions of information quality. These dimensions can be used to make the information quality concept more concrete and measurable. Several studies have confirmed information quality is a multi-dimensional concept (Wang and Strong, 1996). A review of information quality literature reveals a multitude of frameworks which were created in order to investigate information quality within information systems. The most norable of these frameworks is the framework created by Wang and Strong. They formulated fourteen information quality dimensions and grouped them within four information quality categories: intrinsic, contextual, representational, and accessible. Wang and Strong (1996)'s use of dimensions has been adopted by many other researchers, who have refined or finetuned the model to their own research context.

To get a complete view of information quality definitions and definitions of information quality dimensions, eleven other frameworks were inspected to find the most prevalent dimensions of information quality. These dimensions can, in turn, be used to construct the data quality interdepency model with respect to their business impacts.

Table 1 shows the results of the literature research on information quality dimensions. The author and model names, and the quality dimensions used are summarized. A brief selection of the eleven frameworks which were investigated are now described below.

Zeist and Hendriks (1996) identify the information quality characteristics categories functionality, reliability, efficiency, usability, maintainability and portability. The category functionality includes the characteristics suitability, accuracy, interoperability, compliance, security and traceability. Reliability covers the characteristics maturity, recoverability, availability, degradability and fault tolerance. The category efficiency contains the time and resource behaviour. Usability includes the understandability, learnability, operability, luxury, clarity, helpfulness, explicitness, customisability and user-friendliness characteristics of information. Maintainability pertains to the characteristics analysability, changeability, stability, testability, manageability and the reusability. Finally, the category portability contains the characteristics adaptability, conformance, replaceability and installability.

Alexander and Tate (1999) suggest a quality framework for the web and it includes criteria such as authority, accuracy, objectivity, currency, orientation and navigation.

Shanks and Corbitt (1999) describe a semiotic-based framework for the quality of data and it consists of four semiotic levels. Syntactic information quality covers the characteristic consistency. Semantic information quality includes the characteristics accuracy and completeness. The information must be comprehensive, unambiguous, meaningful and correct. Pragmatics information quality include the characteristics usability and usefulness. Furthermore, they list the characteristcs timeliness, conciseness, accessibility and reputation.

Information quality criteria as mentioned by authors Naumann and Rolker (2000) include subject, object and process criteria. Subject criteria cover believability, concise representation and understability of information, interpretability and relevancy of information and added value. Objective criteria include completeness, security, objectiveness, timeliness and verifiability. Process criteria ensure that information should be accurate, have proper linkage to other information, be available, and concise.

Table 1. Data quality dimensions in twelve existing frameworks.

#	Author(s)	Data quality model	Components summary
01	Wang and Strong (1996)	A Conceptual Framework for Data quality	4 categories 16 dimensions
02	Zeist and Hendriks (1996)	Extended ISO Model	6 quality characteristics 32 sub-characteristics
03	Alexander and Tate (1999)	Applying a Quality Framework to Web Environment	6 criteria
04	Katerattanakul and Siau (1999)	IQ of Individual Web Site	4 categories
05	Shanks and Corbitt (1999)	Semiotic-based Framework for Data Quality	4 semiotic descriptions 4 goals of information quality 11 dimensions
06	Dedeke (2000)	Conceptual Framework for measuring IS Quality	5 quality categories 28 dimensions
07	Naumann and Rolker (2000)	Classification of Information Metadata Criteria	3 assessment classes 22 information quality criteria
08	Zhu and Gauch (2000)	Quality metrics for information retrieval on the WWW	6 quality metrics
09	Leung (2001)	Adapted ISO Model for Intranets	6 characteristics 28 dimensions
10	Eppler and Muenzenmayer (2002)	Conceptual Framework for IQ in the website Context	2 'manifestations' 4 quality categories 16 quality dimensions
11	Klein (2002)	<none>	5 information quality dimensions
12	Kahn, Strong and Wang (2002)	Mapping IQ Dimensions into the PSP/IQ model	2 quality types 4 information quality classifications 16 information quality dimensions

It is apparent that there are similarities between the different frameworks, and that there are some characteristics that have been renamed by certain researchers, but may cover the same subject as previously defined characteristics. The characteristics mentioned in previous research were collected and compared in order to find the characteristics most important to this research.

3. Selecting data quality characteristics

The twelve frameworks in Table 1 were compared, and the most common characteristics were abstracted. Table 2 shows which data quality characteristics are present in the frameworks.

Table 2. Data quality characteristics within the twelve frameworks under investigation.

Characteristic	01	02	03	04	05	06	07	08	09	10	11	12	Occurrences
Accuracy	x	x	x	x	x	x	x	x	x	x	x		11
Timeliness	x	x	x		x	x	x	x	x	x	x	x	11
Accessibility	x	x	x	x	x	x	x	x	x	x		x	11
Relevancy	x	x	x	x		x	x	x	x		x		9
Completeness	x			x	x	x	x			x	x	x	8
Objectivity	x		x		x		x	x		x	x		7
Understandability	x	x				x	x		x	x		x	7
Conciseness	x		x	x			x			x		x	6
Consistency	x		x		x	x				x		x	6
Security	x	x					x		x	x			5

The characteristics were selected due to the fact at least half of the investigated frameworks noted these characteristics as an important part of data quality. However, we did include the security characteristic as well, even though it only occurred five times instead of at least six. Furthermore, believability and reputation should be perceived as results of data quality, not characteristics of data quality; these characteristics are therefore omitted from the new model. The rightmost column in Table 2 shows the number of times a data quality characteristic was mentioned.

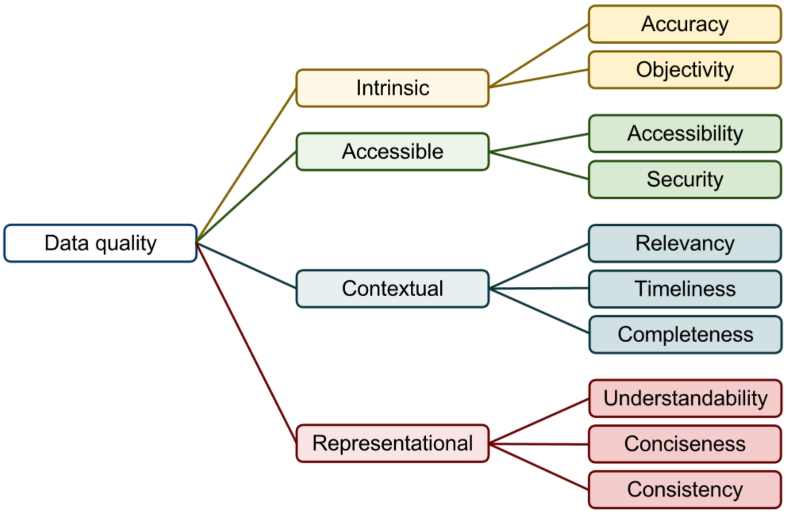

Figure 1. The selection of most frequently occurring data quality dimensions in the literature.

Figure 1 visualizes the selection of most relevant data quality characteristics based on frequency of occurrences in theliterature as this work's condensed data quality model.

4. Conclusions and further research

This research has taken a comparative literature study approach to determine the most relevant data quality dimensions. Relevance in this context has been interpreted as being based on the number of occurrences in existing data quality frameworks within literature. It was found that the most influential framework is the one by Wang and Strong (1996). However, after reviewing eleven other data quality frameworks in the literature, a somewhat different selection of characteristics emerges. This research shows that accuracy, timeliness, and accessibility are, in fact, the Top-3 data quality dimensions, occurring in eleven out of the twelve frameworks under investigation.

Further research will relate this selection of data quality characteristics in Figure 1 to organizational costs and organizational consequences in an effort to better understand their mutual interdependencies.

References

Alexander, J., and Tate, M. (1999). Web Wisdom: How to evaluate and create information quality on the web. Mahwah, NJ.

Dedeke, A. (2000). A conceptual framework for developing quality measures for information systems. In Proceedings if 5th International Conference on Information Quality, pp.126-128.

Eppler, M., and Muenzenmayer, P. (2002). Measuring information quality in the web context: A survey of state-of-the-art instruments and an application methodology. In Proceedings of 7th International Conference on Information Quality, pp.187-196.

Kahn, K., Strong, D., and Wang, R. (2002). Information quality benchmarks: Product and service performance. *Communications of the ACM, 45*, 184-193.

Katerattanakul, P., and Siau, K. (1999). Measuring information quality of web sites: Development of an instrument. In Proceedings of the 20th international conference on Information Systems, pp.279-285, Charlotte, North Carolina, United States.

Klein, B. (2002). When do users detect information quality problems on the world wide web? In American Conference on Information Systems, p. 1101.

Leung, H. (2001). Quality metrics for intranet applications. *Information and Management, 38*(3), 137-152.

Naumann, F., and Rolker, C. (2000). Assessment methods for information quality criteria. In Proceedings of 5th Internation Conference on Information Quality, pp.148-162.

Shanks, G., and Corbitt, B. (1999). Understanding data quality: Social and cultural aspects. In Proceedings of the 10th Australasian Conference on Information Systems.

Wang, R., and Strong, D. (1996). Beyond accuracy: what data quality means to data consumers. *Journal of Management Information Systems, 12*(4), 5-33.

Zeist, R., and Hendriks, P. (1996). Specifying software quality with the extended ISO model. Software Quality Management IV - Improving Quality, BSC, pp.145-160.

Zhu, X., and Gauch, S. (2000). Incorporating quality metrics in centralized/distributed information retrieval on the world wide web. In Proceedings of the 23rd annual international ACM SIGIT conference on Research and development in information retrieval, pp.288-295, Athens, Greece.