Comprehensive Framework for quality Database Applications

A.J.Singh[1], Balvir Singh Thakur[2], Rajesh Chauhan[3], Vikas Sharma[4]

[1, 2 & 3] Department of Computer Science,

[4]ICDEOL

Himachal Pradesh University, Shimla-5 ,

[2]balvir.thakur@gmail.com

Abstract: *The main objective of database analysis, design, and implementation is to establish an electronic repository that faithfully represents the conceptual and logical model of the manageable aspects of a user's information domain. The main goal of data engineering is to put quality data in the hands of user. This paper defines the common dimensions of data quality along with the Framework for database that represents the hierarchy of database quality dimensions including Process, Data factor, Model and behavioral factors.*

Keywords: *Data, Database, Database Quality, Domain knowledge, Information, Quality, Semantic Model.*

Background

While data quality has been the focus of a substantial amount of research, a standard definition does not exist in the literature (Wang & Madnick, 2000). The International Organization for Standardization (ISO) supplies an acceptable definition of data quality using accepted terminology from the quality field. These standards are documented agreements containing technical specifications or other precise criteria to be used consistently as rules, guidelines, or definitions of characteristics to ensure that materials, products, processes, and services are fit for their purpose. Applying the term database quality in this context would build on the ISO

definition of quality, that is, “conformance to requirements” and “fitness for use.” ISO 8402 as a quality management and quality assurance metric provides a formal definition of quality: the characteristics of an entity that represent its ability to satisfy stated and implied needs. This definition is consistent with the notion of customer satisfaction prevalent in the quality literature [1]. Thus, a database can be defined to be of the required quality if it satisfies the requirements stated in a specification, and the specification reflects the implied needs of the user. Therefore, an acceptable level of quality has been achieved if the database conforms to a defined specification, and the specification correctly reflects the intended use. Unfortunately, neither of these definitions is adequate for the purposes of assessing database quality. A database must also be judged by how closely it represents the world of the data consumer (the model), its ability to respond to both routine and unanticipated requests within the domain it is expected to manage (the behavior), and maintain this representation over time. The framework presented herein expands on work previously proposed [2] and incorporates data quality dimensions put forth by several prominent data quality researchers [3]. The framework is important because it expands the definition of strict data quality to that of a broader context of database quality and incorporates the importance of process management.

The Challenge

Many database applications are ultimately unsuitable to the consumer. The process must incorporate three conceptually distinguishable domains: the modeling, the performance, and the enactment domains. Designers attempt to conceptualize the problem domain into a suitable physical model. The proposed physical model is subject to many performance constraints including the physical representation, the network topology, system configuration, and system administration. Finally the most difficult to administer, is the information presented to the consumer for interpretation and enactment. The representation of the database after each of these domain layers all contribute to the quality of the solution by the information consumer. The critical elements below are the bases for the discussion on database quality dimensions.

- The cycle process must be managed toward a successful outcome.

- The model itself must represent a usually diverse and fuzzy problem domain.
- The quality of the data in the database must be of sufficient grade.
- The application must behave or have the ability to behave in a way the consumer understands.

To ensure a quality database application, should the emphasis during model development be on the application of quality-assurance metrics. But there are a significant number of studies and reports that suggest that a large number of database applications fail, are unusable, or contribute to negative organizational consequences (Abate, Diegert, & Allen, 1998; Redman, 1998; Stackpole, 2001; Standish Group, 1997; Wand & Wang, 1996). The Data Warehousing Institute estimates that businesses lose billions each year attributable to bad data (Eckerson, 2002; Trembley, 2002). A quality process does not necessarily lead to a usable database product (Arthur, 1997; Hoxmeier, 1995; Redman, 1995) [4]. There are also many examples of database applications that are in most ways well formed with high data quality but lack semantic or cognitive fidelity (the right design; Motro & Rakov, 1999). Additionally, determining and implementing the proper set of database behaviors can be an elusive task.

While researchers have developed a fairly consistent view of data quality**,** there is little available in the literature on the evaluation of overall database quality including other considerations such as semantic fidelity (model), behavioral, and value factors.

A Database Quality Framework

It is proposed that through the hierarchical framework presented in **Figure 1**,[2] one can consider overall database quality by assessing four primary dimensions: process, data, model, and behavior. Portions of the hierarchy draw heavily from previous studies on data and information quality, and documented process quality standards (Arthur, 1997; Department of Commerce, 2004; Wang, 1998) [5]. A dimension is a set of database quality attributes or components that most data consumers react to in a fairly consistent way (Wang et al., 1996). Wang et al. define data quality dimension as a set of data quality attributes that represent a single data quality abstract or construct. The use of a set of dimensions to represent a quality typology

is consistent with previous quality research (Dvir & Evans, 1996; Strong et al., 1997; Wang et al., 1996). The framework presents the four dimensions in a dimension-attribute-property hierarchy.

Process Quality

Much attention has been given over the years to process quality improvement. ISO-9000-3, total quality management (TQM), quality function deployment (QFD), and the capability maturity model (CMM) are approaches that are concerned primarily with the incorporation of quality management within the process of systems development (Dvir & Evans, 1996; Herbsleb, 1997; Hill, 2003; Schmauch, 1994) [6]. Quality control is a process of ensuring that the database conforms to predefined standards and guidelines using statistical quality measures. Quality assurance attempts to maintain the quality standards in a proactive way. In addition to using quality control measures, quality assurance goals go further by surveying the customers to determine their level of satisfaction with the product. Possibly, potential problems can be detected early in the process.

Database Data Quality

Data integrity is one of the keys to developing a quality database. Atomicity, consistency, isolation, Durability (ACID properties of database) is also used to maintain the quality of data in database. Without accurate data, users will lose confidence in the database or make uninformed decisions. While data integrity can become a problem over time, there are relatively straightforward ways to enforce constraints and domains and to ascertain when problems exist (Moriarty, 1996). The identification, interpretation, and application of business rules, however, present a more difficult challenge for the developer. Rules and policies must be communicated and translated and much of the meaning and intent can be lost in this process.

Data Model Quality

As has been presented, data quality is usually associated with the quality of the data values**.** However, even data that meet all other quality criteria is of little use if they are based on a deficient data model (Levitin & Redman, 1995). Data model quality is the third of the four high-level dimensions

presented above. Information and an application that represent a high proportionate match between the problem and solution domains should be the goal of a database with high semantic quality. Representation, semantics, syntax, and aesthetics are all attributes of model quality[7].

The database design process is largely driven by the requirements and needs of the data consumer, who establishes the boundaries and properties of the problem domain and the requirements of the task. The first step in the process, information discovery, is one of the most difficult, important, and labor-intensive stages of database development [2]. It is in this stage where the semantic requirements are identified, prioritized, and visualized. Requirements can rarely be defined in a serial fashion. Generally, there is significant uncertainty over what these requirements are, and they only become clearer after considerable analysis, discussions with users, and experimentation with prototypes.

Qualitative and quantitative techniques can be used to assist the developer to extract a strong semantic model. However, it is difficult to design a database with high semantic value without significant domain knowledge and experience [8]. These may be the two most important considerations in databases of high semantic quality. In addition, conceptual database design remains more of an art than a science. It takes a high amount of experience, creativity, and vision to design a solution that is robust, usable, and that can stand the test of time.

Database Behavior Quality

Many databases are perceived to be of low quality simply because they are difficult to use. Developers tend to focus on aspects of data quality at the expense of behavioral quality. What constitutes a database of high behavioral quality? Are the criteria different than those used for software applications in general? Clearly the behaviors for a database that is used to support transaction processing (OLTP) are different than those of a database used to support analytical processing (OLAP). Software development, in general, is very procedure or function driven. The objective is to build a system that works (and does it quickly). Database development, on the other hand, should be more focused on the content, context, behavior, semantics, and persistence of the data. The process of behavior

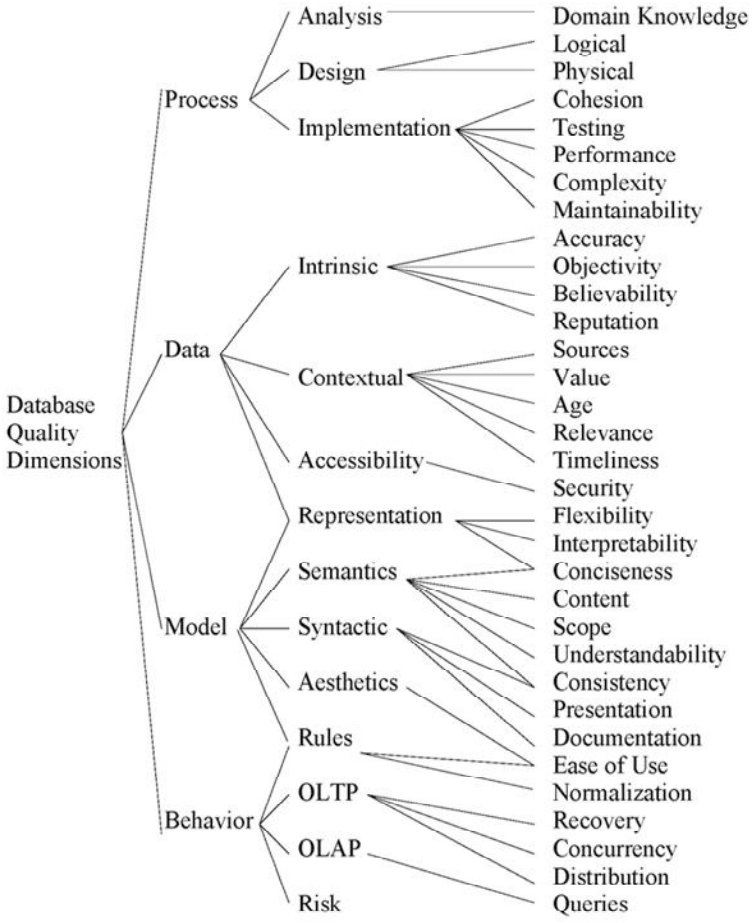
implementation consists of the design and construction of a solution following the identification of the problem domain and the data model.

Table 1: The Common Dimensions of Data Quality

[Wang & Strong; 1996]

	Dimension	Definitions
1	Accuracy	Extent to which data are correct, reliable and certified free of error.
2	Consistency	Extent to which information is presented in the same format and compatible with previous data.
3	Security	Extent to which access to information is restricted appropriately to maintain its security.
4	Timeliness	Extent to which the information is sufficiently up-to-date for the task at hand.
5	Completeness	Extent to which information is not missing and is of sufficient breadth and depth for the task at hand
6	Concise	Extent to which information is compactly represented without being overwhelming (i.e. brief in presentation, yet complete and to the point)
7	Reliability	Extent to which information is correct and reliable
8	Accessibility	Extent to which information is available, or easily and quickly retrievable
9	Availability	Extent to which information is physically accessible.
10	Objectivity	Extent to which information is unbiased, unprejudiced and impartial.
11	Relevancy	Extent to which information is applicable and helpful for the task at hand
12	Useability	Extent to which information is clear and easily used
13	Understandability	Extent to which data are clear without ambiguity and easily comprehended
14	Amount of Data	Extent to which the quantity or volume of available data is appropriate.
15	Believability	Extent to which information is regarded as true and credible.
16	Navigation	Extent to which data are easily found and linked to.
17	Reputation	Extent to which information is highly regarded in terms of source or content
18	Useful	Extent to which information is applicable and helpful for the task at hand.
19	Efficiency	Extent to which data are able to quickly meet the information needs for the task.
20	Value-Added	Extent to which information is beneficial, provides advantages from its use

Figure 1. Database quality dimensions



F

Future Trends

The framework presented above offers a typology for assessing the various dimensions of database quality**.** The purpose of this paper was to expand on the existing research on data and process quality in an attempt to provide a more comprehensive view of database quality. The area is of great concern as information is viewed as a critical organizational asset, and knowledge management and the preservation of organizational memory has become a high priority. It has been estimated by the Data Warehousing Institute, the Gartner Group, Tom Redman, and others that organizations are losing billions of dollars due to poor data quality, and the problem is exacerbated by integrated systems. Most organizations realize that their databases may contain inaccurate data, but they underestimate the business risk of the result: poor information quality (Loshin, 2000). Yet most organizations are unwilling to spend the time and resources necessary to improve the situation. Tom Redman (2004) describes such a scenario in his fictitious case study, Confronting Data Demons. The case study describes a manufacturing firm's struggle to understand and correct data quality problems. Poor data quality is just one aspect of the case. The problem is magnified by the nature of the integrated database (ERP,CRM) and by poorly structured processes.

Further research is required to continue to understand the risks, quantify the costs, improve the model, validate the frameworks, and identify additional data and database quality dimensions.

Conclusion

How does one ensure a final database product that is of high quality? Database quality must be measured in terms of a combination of dimensions including process and behavior quality, data quality, and model fidelity. By organizing attributes into database quality dimensions, many difficulties encountered when dealing with singular attributes can be effectively addressed. So, not only are dimensions more comprehensive, but organizing attributes into dimensions both organizes and minimizes the material that must be comprehended. Moreover, by analyzing dimensions, a data quality researcher may discover systemic root causes of data errors.

References

1. http://www.dtic.mil/ndia/2003CMMI/olson.ppt#256,1,Staged or Continuous: Which Model Should I Choose?
2. John A.Hoxmeier, "A framework for assessing Database quality",Software Quality Journal 7,179-193(1998)
3. "Information Quality Benchmarks: Product and Service Performance", Kahn, B. K., Strong, D. M. and Wang, R. Y., Communications of the ACM, (2002).
4. wikipedia.org/wiki/Data_quality
5. John.A.Hoxmeier "Typology of Database Quality Factors", Software Quality Journal 7,179-193(1998)
6. Shirley Becker "Developing quality complex database systems: practices, Techniques and Technologies "Book by Idea Group Publishing.
7. Defining quality aspects for conceptual models (1995) by J Krogstie, O I Lindland, G Sindre,Proceedings of the Conference on Information System Concepts (ISCO3), Towards a Consolidation of Views, Marburg
8. Denial L.Moody, strategy for improving the Quality of Entity-Relationship Model:A "Toolkit" for practitioners,Department of Information System,University of Melboiurne.
9. Wang,Y.R.&Reddy, M.P.,Kon,H.B.(1993). Toward Quality Data: An attribute Based Approach to appeared in the Journal of Decision Support System(DSS), Special Issue on Information Technologies and System.
10. Ballou, D. and H. Pazer,"Designing information systems to optimize the accuracy timeliness tradeoff", Information Systems Research, Vol. 6, No. 1, 1995, pp. 51-72.
11. Chignell, M. and P. Kamran, Intelligent Database Tools and Applications, Wiley, Los Angeles, California, 1993.
12. Dvir, R. and Evans, S., "A TQM approach to the improvement of information quality", http://wem.mit.edu/tdqm/papers, accessed 7/97.
13. Fox, C., Levitin, A. and T. Redman, "The notion of data and its quality dimensions", Information Processing and Management, 1994, Vol. 30, No. 1, pp. 9-19.

14. Hoxmeier, J. and D. Monachi, “An assessment of database quality: design it right or the right design?”, Proceedings of the Association for Information Systems Annual Meeting, Phoenix, AZ, August, 1996.

15. Levitin, A., and T. Redman, “Quality dimensions of a conceptual view”, Information Processing and Management, 1995, Vol. 31, No 1.

16. Martin, J., Principles of Data-base Management, Prentice-Hall, Inc., Englewood Cliffs, New Jersey, 1976.

17. Moriarty, T., “Barriers to data quality”, Database Programming and Design,, May, 1996, pp. 61.

18. Redman, T.C., “Improve data quality for competitive advantage”, Sloan Management Review, Winter, 1995, Vol. 36, No. 2, pp. 99-107.

19. Rolph, P., and P. Bartram, The Information Agenda: Harnessing Relevant Information in a Changing Business Environment, 1994, London, Management Books 2000, pp. 65-87.

20. Tenopir, C., “Database quality revisited”, Library Journal, 1 October 1990, pp. 64-67.

21. Teorey,T., Database Modeling and Design, The Fundamental Principles, Morgan Kaufman, San Francisco, California, 1994.

22. Wand, Y. and R. Wang, “Anchoring data quality dimensions in ontological foundations”, Total Data Quality.