December 2002

A TYPOLOGY OF DATA WAREHOUSE QUALITY

Cindy LeRouge
University of South Florida

Creggan Gjestland
University of Alabama

A TYPOLOGY OF DATA WAREHOUSE QUALITY

Cindy LeRouge
University of South Florida
clerouge@coba.usf.edu

Creggan Gjestland
University of Alabama
cgjestla@cba.ua.edu

Abstract

Past research has shown that data quality is an important factor in the success of data warehousing projects (Laney, 2000; Wixom & Watson, 2001). This article develops a typology for data warehouse quality that could be used as a guide for future research. Data warehouse quality was divided into information quality and system quality, and attributes for each form of quality were determined based upon a literature review.

Introduction

System quality and information quality are integral antecedents to the potential positive impact a technology can have on an organization (DeLone & McLean, 1992). In following this logic, quality is necessary for successful diffusion of both databases and data warehouses. Though databases and data warehouses are related, these systems cannot be viewed singularly with respect to quality.

This article explores data warehouse quality by contrasting data warehouse and transactional database attributes. A typology is developed from these attributes and then future research is proposed to further develop and refine the typology.

Data Warehouse Quality

Past research has presented a typology of database quality (Hoxmeier, 1998). Though both database and data warehouse technologies include data quality dimensions as a component, the relative importance of certain attributes may differ. To explore the adaptations to existing data quality and database quality typologies necessary to create a data warehouse quality typology, characteristic differences between databases and data warehouses are given in Table 1 and the impact these differences have on data warehouse quality attributes is highlighted.

Delone and McLean (1992) present system quality and information quality as separate exogenous latent constructs in an information system success model. Quality dimensions should be viewed as part of an overall model of success. Therefore, based upon the previous literature review, the following segregated typologies of data warehouse system quality and information quality (Table 2) are presented as a starting point for further investigation of the components of an overall data warehouse system success model. System quality is divided into functionality, usability, and features. Information quality is considered from both input and output perspectives.

Proposed Research

Before using the attributes presented in these typologies as variables in extended research, the comprehensiveness of the model, attribute significance (appropriate inclusion), and relative importance of the factors should be field-tested. Additionally, field-testing may discern contextual or user-based differences in typology components. Structured interviews could be conducted with multiple data warehouse administrators, users and developers to determine which factors are believed to be most important to data warehouse system quality and information quality. These various constituencies have common interests in data warehousing, but provide diversity in perspective.

Table 1. Database and Data Warehouse Comparison

	Database	Data Warehouse	Implications for Data Warehouse Quality Model
Focus	Single purpose data – used as a collection point for transaction and operational data.	Can be used for decision support, knowledge management, and discovery. May be thought of as a business intelligence or expert system to serve analytical purposes.	To be a true knowledge management system, data warehouses systems should manifest growing "intelligence" by capturing aspects of value-added utilization. Predictive ability, perhaps through simulation routines, enhances the quality and value of a data warehouse system.
Interface format	Tailored to transactional processing and transactional summarization.	Tailored to query processing and aggregated information to support decision making	In both systems the user needs to understand the interface. Ease of use promotes utilization and adoption. Clarity is essential for reported information to be interpreted and applied in a data warehouse context, as well.
System Models	Entity/ Relationship	Star Schema, Snowflake Schema. There is cost associated with storing unnecessary data in a data warehouse. Hence, schema elements should be associated with adequate impact paybacks.	To manage costs, the data warehouse should only contain data related to relevant, vital dimensions and avoid redundancy. Additionally, relevant data must be available. Warehouse must be extensible to accommodate "payback" evolution of schemas.
Data sequence	Detail (input transaction) to summary.	Summary to detail (drill down) or further summary (roll up)	It is integral for warehouse data to be flexible, multi-purpose, and capable of assorted aggregation possibilities. Aggregated data should be reconcilable and traceable.
Users	Functional Staff and Management	Mainly senior management	Warehouse data must be trustworthy and accurate. Users may be removed from the operational processes data reflects. Hence, there is no "participatory validation" of results available.
Source data	Current transactional data for daily operations	Appropriate sources are those items required to gain perspective on dimensions (categories and subgroups) of data. May collect the same type of data from a multitude of sources as well as relative data to provide a complete picture.	Data received by the data warehouse must be adequately cleansed and transformed to be useful. Additionally, data correction routines are important (Blackwood, 2000). These functions are not necessary in a database.
Management Purpose	Means of direct control for middle managers. May be thought of as a micro management tool.	Means of organizational control for senior management. May be thought of as a macro management tool.	Data warehouse users require organizational support to take the actions indicated by analyzing system output and other factors.
Soft Data	If not required to complete transaction, not typically considered integral to the system.	Qualitative data is often integral to analyzing various dimensions (Hoxmeier, 1998).	Processing may occur though non-required fields are not complete in a database. These data omissions result in incomplete data for a data warehouse, which can render faulty dimension analysis. Hence, completeness may be more important in a data warehouse.

	Database	Data Warehouse	Implications for Data Warehouse Quality Model
Statistical Analysis	Does not typically provide for robust statistical analysis.	Typically provides for robust statistical analysis either within system or by using supporting software.	To be useful, data warehouses should integrate with supporting analytical software. Hence, it is more critical for data to have quantitative aspects. The resulting analysis must be informative.
Scope	Each database provides for unit or process level analysis.	Affords longititudinal analysis in exploring trends across processes or business units. Can better see the global impact of events.	Information in a data warehouse should be comparable across time.
Truth	Segregated version of truth provided by each database.	Unified version of truth	Data from disparate sources must be successfully integrated within a data warehouse.
Reporting	Conformity in output	Adaptable aggregate output.	Reporting flexibility and adaptability is important in a data warehouse. Data warehouse output should be sufficient for decision-making.
Update method	In congruence with real time operations.	Updated on a scheduled basis. Update patterns from various data sources may be in alternative schedules based upon the timeliness of data needs.	Timeliness in a data warehouse is more closely associated with usefulness and the avoidance of data decay (Blackwood, 2000). An update is timely if changes provide useful information for decision-making.
Performance	Transaction processing time.	Query response time.	Efficiency in a data warehouse system is a balance between performance time and storage costs (particularly of materialized views).
Human Factors	Organizational and system support facilitate database quality.	Organizational support takes on a higher level; a data warehouse should be viewed as a tool for organizational development (Klenz, 1999).	Human factors have increased importance in a data warehouse system. Users and stakeholders should have: • Insight and vision regarding utilization of the technology • The motivation to use the technology • Training on how to use the tool. • Accountability for using the technology

The results of the interviews could be used to develop a survey to further investigate the quality attributes under consideration. The survey may reference a particular data warehouse or have a more global view.

Using tested attributes as factors, quality assurance reviews, compliance audits, user attitude surveys, cost/benefit analysis, budget performance reviews, system performance evaluations, and post installation reviews could be used to assess data warehouse quality and effectiveness for future work. Further analysis may also tease out contextual, task-associated differences in quality perceptions.

Conclusions

Future research utilizing this typology of data warehouse quality could yield results useful to both researchers and practitioners. Validating the proposed typology of data warehouse quality would provide researchers with a basis for exploring one of the factors tied to the success of data warehouse projects. The insights gained from this typology could also be used to better manage data warehouse projects through a clearer understanding of the design and maintenance of data warehouses.

Table 2. A Typology of Data Warehouse Quality

System Quality	
Functional Attributes	• Performance/ Efficiency • Flexibility • Reliability • Extensibility • Maintainability • Timely Updates • Security
Usability Attributes	• Ease of use • Ease of learning • Meets
Feature Attributes	• Knowledge capture • Data cleansing • Data transformation • Powerful statistical analysis • “What if” and simulation • Multi-purpose aggregation • Data integration • Data correction
Information Quality	
Data Input Attributes	• Relevance • Availability • Completeness • Quantifiable • Accuracy • Multi-purpose • Appropriate age • Integratable/aggregatable
Information Output Attributes	• Clarity • Sufficiency • Reconcilable/ Traceable • Valuable • Unique/ Informative • Flexible (roll up/drill down) • Comparable • Accessible (e.g., OLAP, on-line queries) • Multi-format presentation

References

Blackwood, P., Under Construction, Best's Review. 2000: Oldwick, pp. 123-125.

DeLone, W.H. and E.R. McLean. "Information System Success: The Quest for the Dependent Variable," Information Systems Research, 1992. pp. 60-95.

Goldstein, R. "Data Warehousing Solution Focuses on Quality Improvement," Health Management Technology, 2000. pp. 46-48.

Hoxmeier, J.A. "Typology of Database Quality Factors". Software Quality Journal, 1998, pp. 179-193.

Klenz, B.W. The Quality Data Warehouse: Serving The Analytical Needs of the Manufacturing Enterprise, Quality Congress, 1999. Milwaukee: American Society for Quality.

Laney, D. "Data Warehouse Factors to Address for Success," HP Professional, 2000, pp. 21-22.

Wixom, B.H. and Watson, H.J. “An Empirical Investigation of the Factors Affecting Data Warehousing Success,” MIS Quarterly, 2001, pp. 17-41.