Data Quality Problems and Proactive Data Quality Management in Data-Warehouse-Systems

Research Paper

Markus Helfert
University of St. Gallen
Institute of Information Management
Mueller-Friedbergstrasse 8
CH-9000 St. Gallen
(Switzerland)

phone: ++41-71-224 33 82
fax: ++41-71-224 21 89
email: markus.helfert@unisg.ch

Gregor Zellner
University of St. Gallen
Institute of Information Management
Mueller-Friedbergstrasse 8
CH-9000 St. Gallen
(Switzerland)

phone: ++41-71-224 33 48
fax: ++41-71-224 21 89
email: gregor.zellner@unisg.ch

Carlos Sousa
University College Dublin
The Michael Smurfit Graduate School of Business
Dublin
(Ireland)

phone +353-1-716 8811
fax +353-1-716 8993
email: carlos.sousa@ucd.ie

Keywords:

Data Quality, Data Quality Management, Data Warehouse Systems

1 Introduction – Data Quality

Data warehousing has captured the attention of practitioners and researchers for a long period, whereas aspects of data quality is one of the crucial issues in data warehousing [1; 2]. Ensuring high level data quality is one of the most expensive and time-consuming tasks to perform in data warehousing projects [3]. As a consequence of insufficient data quality, frequently data warehouse projects are discontinued [2]. The following article describes major data quality problems, requirements and common strategies to manage data quality in data warehouse systems. The results are based on a survey among large German and Swiss companies, which has been carried out in 2001.

In the following selected approaches for defining data quality are described. Bases on this literature review, an alternative concept will be proposed, which builds the basic framework within this article. Wand and Wang [4] propose a data quality approach focused on the design and operation of information systems. Data quality defects are identified by comparing the information system with the represented part of real world. Each real world state is mapped to a specific state in the information system. Based on this observation, they identify possible representation deficiencies that can occur during system design and data production. These deficiencies are used to define intrinsic data quality dimensions: complete, unambiguous, meaningful and correct. Whereas this approach provides a theoretical base for data quality, it ignores the subjective data quality requirements of the user and concentrates only on the conceptual and internal level.

Another attempt to define and manage data quality is undertaken by English [1]. He identifies different categories for data quality like data definition quality and information architecture quality, data content quality and data presentation quality. For each category a list of detailed quality attributes is proposed. The attributes are overlapping and a description of the numerous connections and dependencies between these attributes is missing. Wang and Strong [5] worked out a list of general data quality characteristics based on a multi-staged empirical survey. The main analysis included 355 questionnaires. Accuracy and correctness were identified as the most important quality attributes for end users. The final results of this study are four categories (Intrinsic, Contextual, Representational, Accessibility) containing several detailed quality attributes. Jarke et al. [6] are proposing a process-oriented classification of the term data quality. Based on this, they link quality factors to the main groups of stakeholders involved in data warehouse projects. Result are prototypical goal hierarchies for each of these user roles [7].

Based on an approach of Garvin [8], an alternative concept for structuring data quality is proposed in this article. According to Garvin, quality approaches can be differentiated into five categories. The *transcendent view* defines quality as a synonym with “innate excellence” or superlative, as a synonym for high standards and requirements. This, rather abstractly philosophical understanding that quality cannot

be precisely defined is insufficient for further work in the context of data quality. Therefore it will not be considered further. *Product-based* definitions are quite different; they view quality as a precise and measurable variable. Quality is so precisely measurable through inherent characteristic of the product. *User-based* approaches start from the opposite premise that quality is stated by the user. Individual consumers are assumed to have different wants, and those products that best satisfy their preferences are those that they regard as having the highest quality. This is a idiosyncratic and personal view of quality, and one that is highly subjective. In contrast to this subjective view, *manufacturing-based* definitions focus on the supply side and are primarily concerned with the production processes. All manufacturing-based definitions virtually identify quality as conformance to requirements. Once a design or a specification has been defined, any deviation implies a reduction in quality. *Value-based* definitions consider terms of costs and prices. According to this view, a quality product is one that provides performance at an acceptable price or conformance at an acceptable cost.

It is important to note, that all these different approaches (disregarding transcendent view) are eligible on different levels of a production system. Each approach serves a special purpose. The different approaches represent the levels of requirement analysis, product and process design and the actual manufacturing process. Therefore they can not be focused separately. Taken this approach for data warehouse systems, three relevant levels for data quality can be identified: The *user-based level* concentrates on the quality demands of the end-user and represents the external level. Starting from these requirements a product specification and a production process can be derived (*product-based, conceptual level*). The product design forms the basis for organizing the manufacturing process (*manufacturing-based, process-oriented level*). From these three quality levels two quality factors can be derived. The factor *quality of design* measures how good the requirements are met by the product design, which is defined in the product specification. *Quality of conformance* compares the final result of the production process with the product specification and gauges the deviation. On the basis of the information requirements the demands for quality of the users will be collected and transformed into a specification e.g., schemes of data bases define entities as well as their properties and thus can be used as a specification for data objects. This specification is the starting point for measuring the quality of the data production process. Quality of conformance looks at the data values and evaluates the compliance with the specification during the data supply process.

2 Data Quality in Data Warehouse Systems

2.1 Research Design

Following this basic data quality framework and based on data quality management [9; 10; 11], a questionnaire of nine questions was developed. Different areas of data quality management (define, measure, analyse, improve) are covered by different questions. The questionnaire was then send through

email to 110 large German and Swiss companies in 2001, whereas 25 completed questionnaires were sent back (respond rate 23%).

The filled in questionnaires point out a broad spectrum of different data warehouse systems (see Figure 1). A concentration of more than 80% on analytical systems as well as report- and controlling systems can be found, whereas a clear classification to system types is not possible. If the filled in questionnaires are analysed by supporting business management functions, the analysis shows as following (see Figure 2): Conspicuous is the high concentration on distributin and sales support.

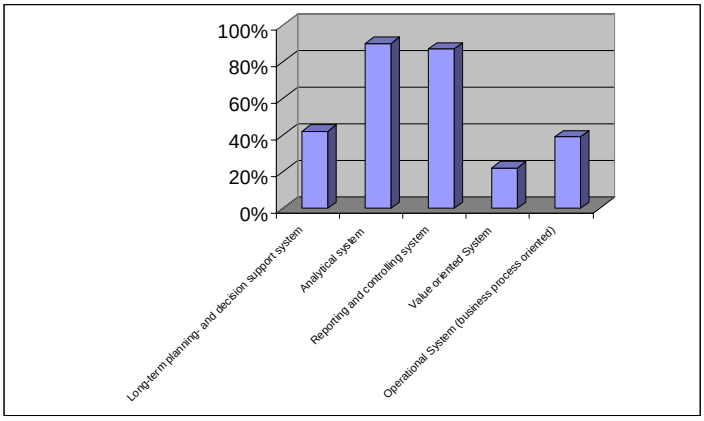

Figure 1: Purpose of Data-Warehouse-Systems[1]

Analysing the scope of business functions, a broad spectrum can be found both in professional aspect and relating to the areas of responsibility (see Figure 3). The concentration on conceptual emphases in the field of the central data warehouse data base and the data marts is striking (see . As expected the specific know-how takes off in direction of operational systems. The technical emphasis lies on the transformation component and the central data warehouse data base. This concludes from the complexity of these transformation processes, which are necessary to map the operational systems to the data warehouse data base. Here are the main problems and causes of integrating different data sources.

[1] Information in % of the filled in questionnaires.

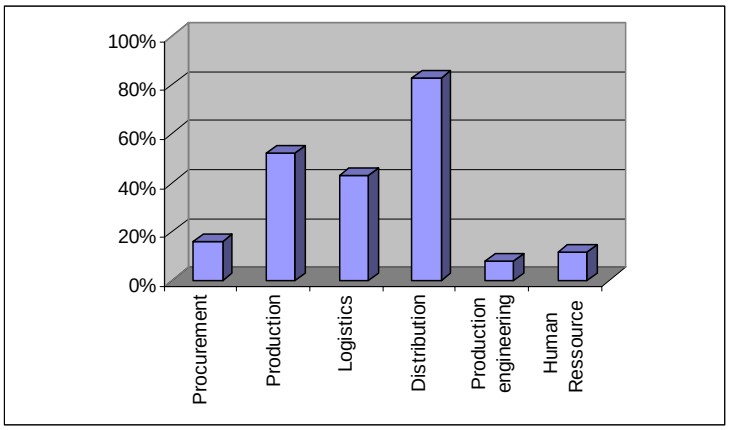

Figure 2: Supported business management functions[2]

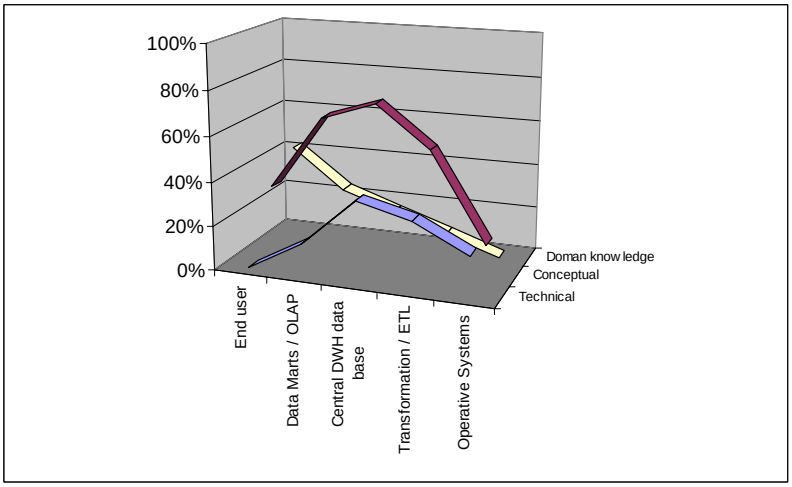

Figure 3: Scope of function of the interviewed persons[3]

2.2 Results and conclusion

If the relevance of the problem area and the measures for ensuring data quality are examined, the outcome represented in Table 1 results. In contrast to 28 % for which data quality does not represent any particular problem, 60 % state data quality as a relevant problem. For 8 % of the responses data quality even represents a high relevancy. For ensuring data quality the majority uses both cleansing techniques and data quality management. Merely 12 % indicate that in their enterprises no special measures are taken.

[2] Information in % of the filled in questionnaires.
[3] Information in % of the filled in questionnaires.

If the measures for ensuring data quality are analysed in detail, basically one can recognise organisational structures apart from technical based data cleansing measures. Organisational units are implemented in the form of data quality officers, business-information-managers, data-owners or data-stewards. However, in the majority of the enterprises the formation of such organisational roles do not seem to be currently completed. Not all enterprises have clear defined responsibilities for data quality. Particularly one can see conflicts between the operative systems and the technical oriented units (IT infrastructure and system developers). For example, communication between different organisational units is necessary for the specification of data quality requirements and the identification of data quality problems.

	Data Cleansing	**DQM**	**DQM and Data Cleansing**	**No special measures**	**Sum**
High relevant	0%	4%	4%	0%	8%
Relevant	4%	20%	36%	0%	60%
No particular problem	8%	4%	4%	12%	28%
No problem	0%	0%	0%	0%	0%
Not specified	4%	0%	0%	0%	4%
Sum	16%	28%	44%	12%	100%

Table 1: Data quality in data warehouse systems[4]

2.2.1 Data quality specifications

Asking questions about quality specification the answers can be classified into two categories. In some enterprises no specification of data quality requirements have taken place. However, these are considered necessary and are often planned already. In particular data quality checks seem to be important. The second category names quality specification through defined standards and delivering agreements.

In these companies

- unambiguously data descriptions,
- formal information about the syntax,
- the delivery time as well as
- specific information about data characteristics (e.g. number of errors and warnings)

are defined. Therefore often historical data and loading processes are analysed and plausibility values are derived.

[4] Information in % of the filled in questionnaires.

2.2.2 Data quality checks

Most quality checks are carried out with plausibility checks. Often, these checks are taken place continuous or frequently at certain dates. An analysis of the quality checks gives the following result:

- 76 % carry out quality checks during the data delivery (Extraction, transforming, loading (ETL) process).
- 68 % examine data quality by the end users during the data usage.
- 60 % analyse data quality of the data warehouse database (Statistical methods, data mining and integrity constrains).
- 52 % ensure data quality within data modelling and system development (standards and organisational measures).
- 20 % check data quality frequently (organisationally regulated).

The following data quality checks are indicated:

- Value ranges and data types (technical check).
- Duplicates by means of key values.
- Referential Integrity.
- Conformance to reference data (e.g. key tables).
- Analysis of system protocols (e.g. refused data records in the ETL process).
- Plausibility (comparison within the data set and within time).
    - Amounts of records and transactions (e.g. number of bookings).
    - Sum of all data records within a relation (e.g. sales) and data distributions.
    - Data verification with other systems and data sources.
- Check through end users (e.g. complaints).
- Through customers (Announcement of data errors through narrow contact to the customer).

Apart from an automatic check that is already partly used, the quality checks are mostly carried out manually with standardised evaluations.

2.2.3 Insufficient Data quality

Essential problems and causes of insufficient data quality are listed as:

- Incorrect values, missing values and references, duplicates (these do not interfere operative systems).
- Inconsistent data between different systems.
- Incorrect data gathering and data operations.
- Insufficient plausibility checks in operative systems (e.g. during data input).

- Changes in the operative systems that are not documented or forwarded to other systems.
- Insufficient modelling and redundant data.
- System problems (technical)

2.2.4 Ensuring data quality

Dealing with data quality problems is diverse and depends on the importance of the problem, the context of usage and the problem cause. Possibilities for ensuring data quality are mentioned as:

- Quality checks before the final loading into the Data Warehouse Data Base.
- Data Cleansing in the ETL process.
- Loading and tagging problematic data (e.g. referential integrity or value domains)
- Automatic correction and data cleansing (e.g. format errors).
- Manual correction and data cleansing (e.g. data interpretation; frequent the problems are already known by the domain expert).
- Feedback to data suppliers about test results (for possible data correction and further data delivery).
- Error location and co-ordination with data suppliers.
- Organisational approaches.

Interestingly not one enterprise listed the integration of the quality specification and quality measurement in the meta data management. If possible the data quality lacks should be reported to the data suppliers and improvement should start at their causes (proactive). A continuous contact between the central data warehouses and source systems is therefore useful.

2.2.5 Data quality characteristics

During the evaluation of data quality itself, the following result have been found. First design quality and quality of conformance are to be distinguished. Design quality is concreted in the information of standards, data definitions and documentation. Quality of conformance focuses on the data contents and the data values. A further important aspect of data quality is quality of the data delivery processes. There, particularly the software components of the overall system have to be considered. The importance of the individual aspects are summarised in Figure 4.

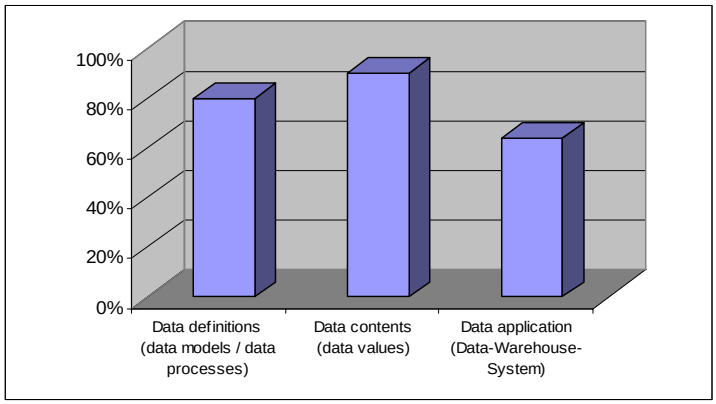

Figure 4: Aspects of data quality[5]

If the characteristics of data quality get analysed, consistency is very important. The database should be consistent with regard to the contents and within time. For data values completeness and correctness as well as the representation of missing values are important. Apart from these, availability, timeliness, referential integrity and syntactic correctness of the data value is important. The tracking of the data source and the documentation of insufficient data quality is relevant. Further semantic and identifiably of data are important for data quality. Here a homogeneous, clear and precise description of the data models and data flows is to be named in particular. The precision of value ranges and the granularity of the data models seem to be less critical. System technical aspects, data protection and access rights are less important and are not considered as characteristics of data quality. Data quality characteristics mentioned by means of an open question are shown in Table 2.

Characteristic	**Sum**	**Characteristic**	**Sum**
Consistent	17	Availability	2
Right, correct, error-free	9	Consistency with value range	2
Complete	9	Accessibility	2
Clear defined (data`s characteristic qualities, interfaces, documentation)	7	Data quality Management Concept	1
Up to date	6	Conformance to experiences (domain expert)	1
Standardised form	4	Conformance to source data	1
Time-related	4	Availability	1
Interpretability / Understandability	3	No dublicates	1
Traceability (data source)/ variations documented	3	No bad data	1
Syntactic correctness	3	DQ-officer Motivation	1
Representation of enterprise	2	Timeliness	1
User-related / decision-supporting	2	Security	1
Ability for management analysis	2	Monitored	1
Granularity	2	Comparable	1
Sufficiently precise	2	Centralised	1

[5] Information in % of the filled in questionnaires.

Referential integrity	2	Reliable	1
Relevance	2		

Table 2: Mentioned data quality features[6]

2.3 Conclusion and towards proactive Data Quality Management

Based on the results of the survey an approach for managing data quality in data warehouse systems will be proposed in following. In literature there are several approaches for managing and defining data quality [e.g.1, 2, 7, 12, 13, 14], but still the question remains how to ensure high level data quality in data warehouse systems. To provide a management concept for ensuring high level data quality, current research at the Institute of Information Management applies the concept of total quality management (TQM) to data warehouse systems. Typical for TQM is the orientation on customer requirements, the participation of all stakeholders, continuous improvement and the comprehensive management approach [15, 16]. All enterprise wide activities are integrated into an enterprise wide structure aiming to improve products, services and process quality continuously and therefore satisfying customer requirements. Based on TQM the proposed proactive data quality management enfolds organizational structures with roles and responsibilities as well as quality processes for ensuring continuous quality improvement. The processes are supported by techniques and tools. Standards and guidelines are ensuring a consistent design and operation of the data warehouse system [10].

3 Acknowledgement

The results presented in this paper are part of current research at the Institute of Information Management, University of St. Gallen (IWI-HSG), Switzerland (http://www.iwi.unisg.ch/).

4 References

[1] L. English: Improving Data Warehouse and Business Information Quality. Wiley, New York et al. 1999.

[2] M. Helfert: Massnahmen und Konzepte zur Sicherung der Datenqualität, in Jung, R., Winter, R. (ed.): Data-Warehousing-Strategie: Erfahrungen, Methoden, Visionen, Springer, Berlin et al., 2000, p. 61-77.

[3] C. Haeussler: Datenqualitaet, in Martin, W. (ed.): Data Warehousing, ITP GmbH, Bonn, 1998, p. 75-89.

[4] Y. Wand, R. Y. Wang: Anchoring Data Quality Dimensions in Ontological Foundations in: Communications of the ACM, 39(1996), No. 11, p. 86-95.

[6] Number of nominations; similar statements summarised in groups

[5] R. Y. Wang, D. M. Strong: Beyond Accuracy: What Data Quality Means to Data Consumers, in Journal of Management of Information Systems 12(1996), No. 4, p. 5-33.
[6] M. Jarke, M. Jeusfeld, C. Quix, P. Vassiliadis: Architecture and Quality in Data Warehouses: An Extended Repository Approach, in Information Systems, 24(1999), No. 3, p. 229-253.
[7] M. Jarke, M. Lenzerini, Y. Vassiliou, P. Vassiliadis: Fundamentals of data warehouses, Springer, Berlin et al., 2000.
[8] D. A. Garvin: What does 'Product Quality' really mean?, in Sloan Management Review, 26(1984), No. 1, p. 25-43.
[9] M. Helfert, R. Radon: An Approach for Information Quality measurement in Data Warehousing, in B. D. Klein,.D. F. Rossin (ed.): Proceedings of the 2000 Conference on Information Quality, Cambridge, MA 2000, pp. 109-125.
[10] M. Helfert: Managing and Measuring Data Quality in Data Warehousing, in Proceedings of the World Multiconference on Systemics, Cybernetics and Informatics, Florida, Orlando, 2001, p. 55-65.
[11] M. Helfert, E. von Maur: A Strategy for Managing Data Quality in Data Warehouse Systems, in E. M. Pierce, R. Kaatz-Haas (ed.): Proceedings of the Sixth International Conference on Information Quality, Cambridge, MA 2001, pp. 62-76
[12] K. Huang, Y. Lee, R. Wang: Quality Information and Knowledge, Prentice Hall, Upper Saddle River, NJ, 1999.
[13] G. K. Tayi, D. Ballou: Examining Data Quality, in Communication of the ACM 41(1998), February, No. 2, p. 54-57.
[14] M. J. Eppler, D. Wittig: Conceptualizing Information Quality: A Review of Information Quality Frameworks from the Last Ten Years, in Klein, B. D., Rossin, D. F. (ed.): Proceedings of the 2000 Conference on Information Quality, Cambridge, MA, 2000, p. 83-96.
[15] J. M. Juran: How to think about Quality, in Juran, J. M., Godfrey, A. B. (ed.): Jurans's Quality Handbook, McGraw Hill, Ney York et al., 1999, p. 1-18.
[16] H. D. Seghezzi: Integriertes Qualitätsmanagement – das St. Galler Konzept, Hanser, Munich and Vienna, 1996.

5 Biographies:

Markus Helfert: After gaining a Bachelor of Science at the Napier University (Edinburgh; UK-Scotland) in computer science and business, he finished his studies at the University of Mannheim (Germany) in information management. Then he worked as a research assistant at the Institute for Management and Consulting in Mannheim (Germany), where he participated in several projects in logistics and information management. During a research project at Artificial Life, Inc. (Boston, USA) he worked in the field of data mining, artificial intelligence and e-commerce. Since 1999 Markus Helfert works at the Institute of Information Management, University of St.Gallen (Switzerland) in the field of data warehouse

systems. For his PhD-thesis, his research was focussed on data warehouse systems and data quality management. He has published several articles in the field of data quality management.

Gregor Zellner studied business economics at the University of Munich (LMU, Germany) and the University of Ingolstadt (Germany) and finished his studies in 1999. Afterwards he worked as a research assistant at the institute for economic and social politics (University of Ingolstadt, Germany) in a research project about regional economics. Since July 2000 Gregor Zellner is research assistant at the Institute of Information Management, University of St. Gallen (Switzerland). In 2000/2001 he worked in the centre of competence called “Banking Architectures of the Informationage” focussing business processes and the process layer in business engineering. Currently he works on his PhD thesis, which focuses on defining new CRM-processes, building a process-landscape for relationship management and creating a method to transform theoretical customer-lifetime-value-analysis into a applicable proceeding for the praxis.

Carlos Manuel Sousa is a doctoral candidate in Marketing at the Michael Smurfit Graduate School of Business, University College Dublin, Ireland. His educational background includes a B.A. in Accounting and Management from the Coimbra Institute of Accountancy and Management, Portugal, an Master in International Business from University of Santiago de Compostela, Spain, and an MComm from University College Dublin, Ireland. He has recently published in the Irish Journal of Management.