

# Managing and Measuring Data Quality in Data Warehousing

Markus Helfert

Institute of Information Management, University of St. Gallen

St. Gallen, CH-9000, Switzerland

Markus.Helfert@unisg.ch

## Abstract

High level data quality and the management of ensuring data quality is one of the key success factors for Data Warehousing projects. The following article describes an approach for Data Quality Management, which is based on theories as well as practical experiences. Starting from effects of insufficient data quality in practice, a definition for information, data and data quality will be worked out. Based on the concept of total data quality management the method-based Data Quality Management (DQM) will be described. As key part of method-based DQM an approach for planing and measuring data quality will be illustrated and explained. Finally, based on the research results further conclusions are summarised.

**Keywords:** Data Warehousing, Data Quality, Data Quality Management, Data Quality Measuring, Information Systems

## 1. Effects of insufficient Data Quality

The following paragraph summarizes results of several workshops held within the Competence Center 'Data Warehousing Strategy' (CC DWS) and shows main effects of insufficient data quality in practice. Effects of insufficient data quality can be classified in effects on Data Warehousing projects, effects on decision processes and effects on operative processes.

Expensive search for the 'right and correct' values, additional efforts to create reports and analysis, multiple acquisitions of data, special transformation logic and data cleansing algorithms as well as additional efforts of designing and operating the Data Warehouse are only a few effects on Data Warehousing projects, which are resulting in additional efforts and therefore in higher project costs. Because the Data Warehouse will be used by only a few specialist, it will not be accepted broadly within the company. Furthermore insufficient data quality results in reduced internal acceptance of the Data Warehousing project by operating not reliable and not authentic. The expected benefits will not be achieved and the internal project support could be lost.

The Data Warehouse in general is used to create analysis and reports for supporting decision processes.[22] In practice, often incorrect, meaningless and incomprehensible statistics, analysis and reports exists. These insufficient analytical information lead to deficient decisions and lack of external reporting (e.g. accounting). Deficient decisions based on insufficient information could result in inadequate strategies, accumulation of risks and an inadequate customer orientation with an inadequate pricing policy.

Effects on operative processes are multiple. For example, insufficient (e. g. inaccurate) information used for customer relationship management and electronic commerce can dissatisfy customers, and so resulting in complaints or lost customers. With falsely predicted customer behavior selling potentials are not identified and wrong customer target groups are advertised by inadequate promotions. Cross selling potentials are misinterpreted or even not identified. Incorrect accounts and falsely calculated bonuses resulting in reduced external image and a bad reputation, which also leads to a narrow turnover.

Showing these few examples of insufficient data quality it is not surprisingly that data quality is identified as one of the major problems of Data Warehousing. A survey within large German and Swiss companies confirmed this result [8], which leads us to the development of the method-based Data Quality Management (DQM).

## 2. Data Quality in Data Warehousing

### 2.1. Data Quality Definition

Discussion in literature about information, data and data quality shows that these terms are complex and still no widely accepted definition exists. There are numerous approaches for defining information [3], quality [12] and data quality respective information quality [11, 21, 23] and therefore it is at least necessary to clarify these terms used. Many approaches do not distinguish between data and information and define data quality and information quality equal.

In the following a suitable definition for knowledge, information and data will be described on the basis of five dimensions used in main approaches in economic literature.[3] The dimensions are semiotic, medium, novelty, verity and timeliness as shown in Figure 1.

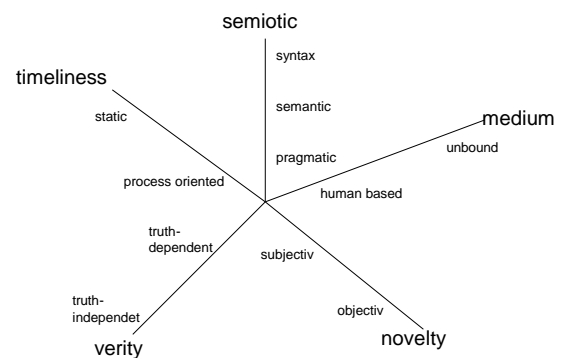


Figure 1: Dimensions of information [3]

Human based and unbound approaches are distinguished by the dimension medium. In human based approaches information exists only within humans whereas unbound approaches considering all mediums (e.g. books and hard discs). Indeed the cognitive interpretation process and the matching to real world objects can only be done by humans, but the expressed information by itself has already a specified or intended meaning. For this reason, the article comprises as information all representations, which can be stored and transferred via some medium.

Information in approaches based on subjective novelty have to be new for information consumers whereas in inter-subjective or objective approaches information is independent from the subject and its previous knowledge. Because the character of information has to be verified in each individual case, subjective approaches tend to be less operational than objective ones. An objective approach define information based on some characteristics and is independent from single cases. Therefore these approaches are more qualified for the purpose of measuring information and data quality than subjective ones.

Based on verity approaches can be distinguished in truth-dependent and truth-independent approaches. In truth-dependent approaches information has to be true whereas in truth-independent approaches verity is irrelevant.

There are static and process oriented approaches, which are forming the dimension timeliness. Process oriented approaches considering information as process of information acquiring. On the other hand, static approaches define information as a state, as input and result of information processes.

Semiotics provides a framework consisting of three levels: syntax, semantic and pragmatic.[3, 25] The bottom level of the framework, the syntax, considers basic representation of information. This level deals with technical context and syntactical concerns and comprises characters, symbols and signals. These basic representation are summarized by the term signals, which are comprising all symbolic representation regardless of their meaning. The next level, the semantic level, deals with related real world objects and implies some meaning to signals. The pragmatic level, as the top level, deals with information processes and information users. The information usage has some purpose and some impact to the information process and the information user. The framework provided by semiotics gives a basic structure for different levels of information. Every level includes their previous levels, whereas the transition of levels is fuzzy and subjective.

Information in this article should be defined as knowledge, which can be expressed in some human language, whereas knowledge represents some real world objects. Information defined as this, is a subset of knowledge. This definition of information is characterized by the five dimensions as unbound, objective, truth-independent, static and on the semantic level. Following this definition, data can be defined as a subset of information, which can be processed by machines.

The term quality is as complex as the term information.[12] One approach, which is widely adopted in quality literature, is focused on the consumer and the product's fitness for use. This approach comprises two aspects of quality. First, quality means product characteristics, which meet customer needs and thereby provide customer satisfaction and second absence from deficiencies that result in customer dissatisfaction.[12] The first aspect refers to quality of design and the second to quality of

conformance. Quality of design addresses the aspect of information requirements and information product design. How good are the requirements met by the information product design? The conformance of the final information product with the product design is addressed by quality of conformance. Quality of conformance take the divergence of design with the final product into consideration. Because low quality of design and low quality of conformance have different causes and therefore different solutions, it should be considered differently. High quality of design do not mean high quality of conformance and vice versa. Increasing quality of design tends to result in higher costs, whereas increasing in quality of conformance tends to results in lower costs. In addition, higher conformance means fewer complains and therefore increased customer satisfaction. Therefore data quality measurement has to consider both aspects of quality.

## 2.2. Method-based Data Quality Management

Quality management includes concepts of quality policy, quality planning, quality control and quality assurance as well as quality improvement. Quality management operates through the quality management system with organizational structure, process organization, standards, guidelines, rules, methods, techniques, and tools as their elements.[19, 20]

One widely accepted concept for quality management is the concept of total quality management (TQM). The concept states the current research in quality management and is already successfully implemented in manufacturing sector. Currently the concept of TQM is applied to other sectors like service industry and data quality.[5, 17, 25] Typical for TQM is the orientation on customer requirements, the participation of people, continuous improvement and the comprehensive management approach. All enterprise wide activities are integrated into an enterprise wide structure aiming continuously improvement of product, service and process quality and therefore satisfy customer requirements.[20]

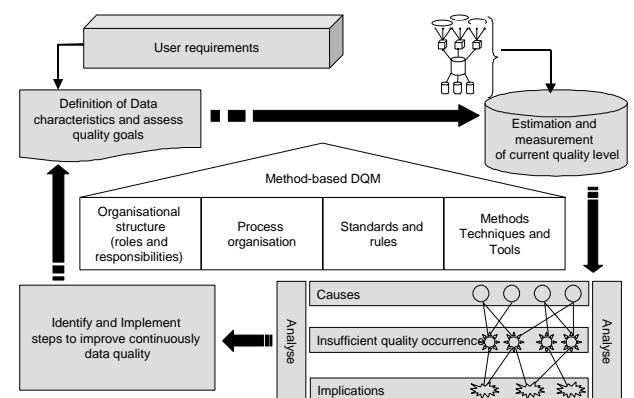


Figure 2: Method- based Data Quality Management [9]

Current research within the Competence Center 'Data Warehousing 2' applies the concept of Total Quality Management to Data Warehousing and in particular to data quality. First results showing that concepts, principles and techniques applied in manufacturing and service industry can be transferred with some adaptations to Data Warehousing.[10] Based on TQM, the proposed method-based Data Quality Management (DQM) [9] enfolds organizational structure with roles and responsibilities.

The process cycle ‘Define’, ‘Measure’, ‘Analyse’, and ‘Improve’ is building the core process organization for ensuring continuous quality improvement. The processes are supported by methods, techniques, tools, standards, guidelines and rules, which are based on method engineering [7] and providing therefore an integrated framework. Figure 2 shows the method-based DQM approach as it is currently developed.

Before analyzing and improving data quality it is essential to plan, define and assess quality goals and measure current quality levels. Data quality planning and measuring are therefore key success factors of the proposed data quality management concept. Data quality characteristics and their appropriate measuring techniques building a data quality framework and giving the possibility to state current quality levels. By the comparison in time, it is possible to identify quality trends and evaluates the effects of quality improvements. The framework also provides the foundation for cost benefit analysis and for quality improvements. In following an approach for a suitable data quality framework in Data Warehousing is presented.

### 2.3. Data Quality Framework

In management and information technology literature there are currently many information quality frameworks and approaches for measuring data quality.[6] Table 1 summarizes a selection of information quality frameworks from various contexts. Besides these frameworks there are a large number of information quality criteria lists. An evaluation of these frameworks and criteria lists shows that there is still lack of quality indicators and measurement systems. On the one hand, data quality can be measured with subjective perceptions from information users. On the other hand, there are approaches developing measuring systems on the basis of quality characteristics (mostly intrinsic information quality characteristics like for example completeness and correctness).[14] But as of today no integrated data quality framework with a generic, generally applicable measuring system is available yet. Most frameworks only provide limited assistance for analyzing causes of insufficient quality as well as providing guidelines for solving identified problems. Furthermore the most frameworks lack of providing methods to apply the framework to company specific requirements.

Author and year of Publication	Application context
Augustin/Reminger 1990	Management Information Systems
Morris et al. 1996	Management
Redmann 1996	Data Bases
Miller 1996	Information Systems
Wang/Strong 1996	Data Bases
Davenport 1997	Information Management
Ballou et al.1998	Data Warehousing
Kahn/Strong 1998	Information Systems
Rittberger 1999	Information Service Providers
English 1999	Data Warehousing
Huang et al. 1999	Knowledge Management

Table 1: Information quality frameworks [6]

As an example of an information quality framework Huang et al. constructing, based on a customer oriented quality approach and empirical studies a framework in the context of knowledge management with four information quality categories and related information quality criteria.[11]

IQ Category	IQ Criteria
Intrinsic	Accuracy, objectivity, believability, reputation
Contextual	Relevancy, value-added, timeliness, completeness, amount of information
Representational	Interpretability, ease of understanding, concise representation, consistent representation
Accessibility	Access, security

Table 2: Example of an information quality criteria list [11]

The framework is of empirical relevance and gives a suitable foundation for further research, but it is defined on a high level and is not explicit integrated within the proposed measuring system. As shown above, it is necessary to differ between quality of design and quality of conformance on different semiotic levels. The proposed framework should define precisely related data quality criteria, show relationships between them and how each quality criteria contribute to data quality. Finally, a measuring system with techniques for assigning values to quality criteria should be integrated within the framework.

Huang et al. are suggesting a measuring system for information quality consisting of three metrics. First, a metric that measures individual’s subjective estimates of information quality is suggested. The metric consists of a questionnaire with simple and for the information user understandable questions. The information user is questioned about his or her subjective estimation of information quality in their context of information usage. Second, a metric that measures information quality along quantifiable, objective variables that are application independent is proposed. The metric is basically based on established theory for controlling the quality of data entering. Examples of objective variables are given as correctness, completeness and consistency. Third, a metric that measures information quality along quantifiable, objective variables that are application-dependent, is suggested. This third measurement needs knowledge about information, their application and their formats.

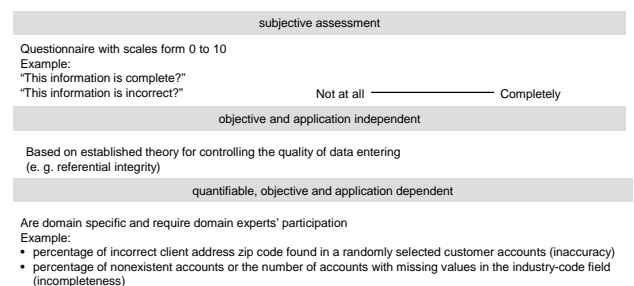


Figure 3: Measuring system proposed by Huang et al.

Even if the proposed measuring system, which is summarised in Figure 3, applies different metrics for measuring information quality, further research have to be done. The description is still

based on few examples and therefore a method for adapting the metrics to specific situations should be developed. Furthermore the technical representation should be described and their relations to data quality criteria should be shown. The different metrics are overlapping so that a clarification of the metric's focus is necessary.

In following a modified data quality criteria list and an approach for integrating measuring techniques is shown. The approach is mainly based on results of workshops held within the competence center and building the foundation for further research. The results are focused particularly on measuring data quality in Data Warehouse Systems and — as part of the method-based DQM — showing research areas for developing a comprehensive data quality framework.

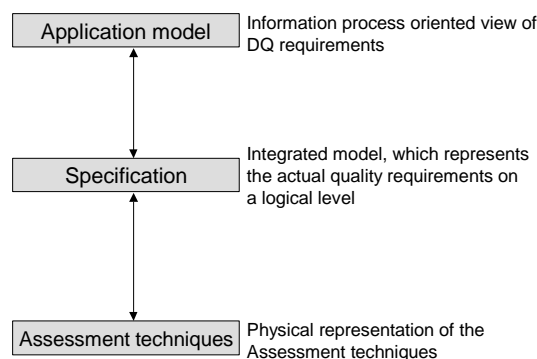
Core element of the framework is a set of relevant quality characteristics. These characteristics are classified on the three levels of semiotics and the two aspects of quality. At the pragmatic level all characteristics, which are relevant within the application of data for information processes, are located. For example, relevance, completeness are characteristics for quality of design and timeliness, actuality and efficiency are important characteristics for quality of conformance. At the semantical level, which deals with the meaning of data, interpretability, accuracy as well as consistent and complete data values, believability and reliability are important for quality of conformance. Quality of design on the semantic level is characterized by precise as well as easy to understand and objective data definitions. Quality of design on the level of syntax comprise consistent and adequate syntax. Syntactical correctness, consistent representation, security and accessibility are characteristics of quality of conformance. Table 3 summarizes the proposed data quality characteristics.

Semiotic Level	Quality Characteristics		Measurement Approach
	Quality of Design	Quality of Conformance	
Pragmatic	Relevance, completeness	Timeliness, actuality, efficiency	Information process, application
Semantic	Precise data definitions, easy to understand and objective data definitions.	Interpretability, accuracy, consistent data values, complete data values, believability, reliability	Comparison with real world and experience
Syntax	Consistent and adequate syntax	Syntactical correctness, consistent representation, security, accessibility	Syntactical standards and agreements

**Table 3: Data quality characteristics**

Data quality characteristics are founding the core element for planing and measuring data quality in Data Warehousing. Figure 4 shows an integrated approach for planing and measuring data quality. The top level states the subjective quality requirements based on information processes and information users. For the communication with information user a conceptual modeling language is necessary. The next level represents an integrated model, which represents the specified quality requirements on a logical level. The specification of quality requirements, which is based on the proposed data quality

characteristics, objectify the quality measures. For assigning values to these quality characteristics techniques based on statistics (e. g. statistical quality control) and data mining as well as questionnaires are applicable. On the bottom level physical representation of assessment techniques relates the measurement system and assigning techniques to the quality specification.



**Figure 4: Data quality framework**

### 3. Conclusions

The proposed data quality framework provides an integrated approach for planing and measuring data quality in Data Warehousing. The framework builds, as key part of the method-based Data Quality Management, the base for ensuring high level data quality in Data Warehouse Systems. Even if this article can not provide a detailed description, the finding shows areas of further research and provides a way to consolidate further results into an integrated framework. Further research has to be done in the areas of

- assessment techniques (e. g. applying data mining and statistics to data quality),
- developing a generic quality model and a method for applying this model to specific situations,
- relating assessment techniques to quality criteria as well as
- modeling data quality requirements.

Current research at the competence center 'Data Warehousing 2' focuses on modeling data quality requirements as well as evaluating possible assessment techniques and integrates these techniques to a consisted and integrated measuring system.

### 4. Acknowledgements

The Competence Center 'Data Warehousing Strategy' CC DWS, was founded at the University of St. Gallen, Switzerland, in January 1999. The CC DWS was a joint research project of the Institute of Information Management and 12 large German and Swiss companies from insurance, logistics, telecommunications, banking and consulting industry, and the Swiss department of defense. The successor of the CC DWS continues the research in Data Warehousing and is named Competence Center 'Data Warehousing 2'. (<http://datawarehouse.iwi.unisg.ch>)

## 5. References

- [1] Augustin, S., Reminger, B.: Trotz Datenflut jede Menge Informationsdefizit! – Ist das erfolgreiche JIT-Konzept auch in der Info-Welt realisierbar?, in Baeck, H. (ed.): Der informierte Manager, Koeln: TUEV Rheinland, 1990, pp. 73-83.
- [2] Ballou, D. P., Wang R., Pazer, H., Tayi, G. K.: Modelling information manufacturing systems to determine information product quality, in Management Science, April 1998, Vol. 44, Issue 4, pp. 462-484.
- [3] Bode, J.: Der Informationsbegriff in der Betriebswirtschaftslehre, in Schmalenbachs Zeitschrift für betriebswirtschaftliche Forschung, 1997, Vol. 49, Issue 5, pp. 449-468.
- [4] Davenport, T.: Information Ecology: Mastering the Information and Knowledge Environment, Oxford: Oxford University Press, 1997.
- [5] English, L.: Improving Data Warehouse and Business Information Quality, New York: Wiley & Sons, 1999.
- [6] Eppler, M. J., Wittig, D.: 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: Massachusetts Institute of Technology, 2000, pp. 83-96.
- [7] Gutzwiller, T. A.: Das CC-RIM-Referenzmodell für den Entwurf von betrieblichen, transaktionsorientierten Informationssystemen, Heidelberg: Physica-Verlag, 1994.
- [8] Helfert, M.: Eine empirische Untersuchung von Forschungsfragen beim Data Warehousing aus Sicht der Unternehmenspraxis, Working Paper BE HSG/CC DWS/05, Institute of Information Management, University of St. Gallen, 2000.
- [9] Helfert, M.: Massnahmen und Konzepte zur Sicherung der Datenqualität, in Jung, R.; Winter, R. (ed.): Data Warehousing Strategie – Erfahrungen, Methoden, Visionen – Berlin et al.: Springer, 2000.
- [10] Helfert, M., Radon, R.: An Approach for Information Quality measurement in Data Warehousing, in Klein, B. D., Rossin, D. F. (ed.): Proceedings of the 2000 Conference on Information Quality, Cambridge, MA: Massachusetts Institute of Technology, 2000, pp. 109-125.
- [11] Huang, J., Lee Y. W., Wang R. Y.: Quality Information and Knowledge; Upper Saddle River, NJ: Prentice Hall 1999.
- [12] Juran, J. M.: How to think about Quality, in Juran, J. M., Godfrey A. B. (ed.): Juran's quality handbook, 5<sup>th</sup> ed., New York: McGraw-Hill, 1998, pp. 2.1-2.18.
- [13] Kahn, B. K.; Strong, D. M.: Product and Service Performance Model for Information Quality: An Update 1998, in Chengalur-Smith, I., Pipino, L. L. (ed.): Proceedings of the 1998 Conference on Information Quality, Cambridge, MA: Massachusetts Institute of Technology, 1998.
- [14] Kaomea, P.: Valuation of Data Quality – A Decision Analysis Approach, Working Paper TDQM-94-09, Sloan School of Management – Massachusetts Institute of Technology, Cambridge, MA, 1994.
- [15] Miller, H.: The multiple dimensions of information quality, in Information Systems Management, Vol. 13, Issue 2, Spring 1996, pp. 79-82.
- [16] Morris, S., Meed, J., Svensen, N.: The Intelligent Manager, London: Pitman Publishing, 1996.
- [17] Redman, T. C.: Data quality for the information age, Norwood: Artech House, 1996.
- [18] Rittberger, M.: Certification of Information Services, in Lee, Y. W., Tayi, G. K. (ed.): Proceedings of the 1999 Conference on Information Quality, Cambridge, MA: Massachusetts Institute of Technology, 1999, pp. 17-37.
- [19] Schwarze, J.: Informationsmanagement – Planung, Steuerung, Koordination und Kontrolle der Informationsversorgung im Unternehmen, Herne, Berlin: Neue Wirtschafts-Briefe, 1998.
- [20] Seghezzi, H. D.: Integriertes Qualitätsmanagement: das St. Galler Konzept, Muenchen, Wien: Hanser, 1996.
- [21] Tayi, G. K., Ballou D.: Examining Data Quality, in Communications of the ACM, 1998, Vol. 41, Issue 2, pp. 54-57.
- [22] von Maur, E.: Object Warehouse - Konzeption der Basis objektorientierter Management Support Systems am Beispiel von Smalltalk und dem ERP Baan, Osnabrück, Univ., Diss., 2000.
- [23] Wand, Y., Wang R.: Anchoring Data Quality Dimensions in Ontological Foundations, in Communications of the ACM, 1996, Vol. 39, Issue 11, pp. 86-95.
- [24] Wang, R. Y., Strong, D. M.: Beyond Accuracy: What Data Quality Means to Data Consumers, in Journal of Management Information Systems, Spring 1996, Vol. 12, Issue 4, pp. 5-33.
- [25] Wolf, P.: Konzept eines TQM-basierten Regelkreismodells für ein „Information Quality Management“ (IQM), Dortmund: Praxiswissen, 1999.