Proactive Data Quality Management
for Data Warehouse Systems

- A Metadata based Data Quality System -

Markus Helfert
University of St. Gallen
Institute of Information Management
markus.helfert@unisg.ch

Clemens Herrmann
University of St. Gallen
Institute of Information Management
clemens.herrmann@unisg.ch

Abstract

Data warehousing has captured the attention of practitioners and researchers for a long time, whereas aspects of data quality is one of the crucial issues in data warehousing [ENG99; HEL00]. Still, ensuring high level data quality is one of the most expensive and time-consuming tasks to perform in data warehousing projects [HAE98]. Many data warehouse projects are discontinued due to insufficient data quality [HEL00]. The following article describes an approach for managing data quality in data warehouse systems through a metadata based data quality system. The results are integrated in a comprehensive management approach and are based on practical experiences within a Swiss bank.

1 Data Quality Framework

1.1 Proactive Data Quality Management

There are several textbook approaches for managing and defining data quality [e.g. JAR00; HEL00; HUA99; ENG99; TAY98, EPP00], 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 by the Competence Center 'Data Warehousing 2'[1] applies the concept of total quality management (TQM) to data warehouse systems. Focus on customer requirements, participation of all stakeholders as well as continuous improvement and a comprehensive management approach are important characteristics of TQM [JUR79; SEG96]. All enterprise wide activities are integrated into an enterprise wide structure which continuously improves products, services and process quality in order to satisfy customer requirements. Proactive data quality management (proDQM) is based on TQM and combines an organizational structure defining functions and responsibilities with processes ensuring continuous quality improvement. Techniques and tools support the processes, standards and

[1] The results presented in this paper are part of the Competence Center 'Data Warehousing 2' (CC DW2) of the Institute of Information Management, University of St. Gallen (IWI-HSG), Switzerland, http://datawarehouse.iwi.unisg.ch

guidelines ensure consistent design and operation throughout the data warehouse system [HEL01]. proDQM is based on method engineering proposed by [GUT94] and thus provides a comprehensive method for data quality management in data warehouse systems.

The operational level of proDQM consists of the following two key tasks:

- In *quality planning* data consumer requirements and expectations are gathered and then transferred into data delivery processes and specifications. Quality criteria are selected, classified and prioritized in the process.
- *Quality control* verifies the data delivery processes and assures they comply to the specifications. In order to identify and implement the adequate steps, data quality must be measured quantitatively.

Because of the importance of data quality planning and control, the current article focuses on these tasks. It is essential to plan, define and assess quality goals and measure current quality levels. In the following, some related approaches for defining and measuring data quality are sketched. Then a possible essential classification for data quality which provides the basic conceptual framework for a real world metadata based data quality system is introduced. After this description of the basic architecture some practical applications are discussed.

1.2 Data Quality

[WAN96a] introduces a data quality control 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 the real world. Each real world state is mapped to a specific state in the information system. Based on this observation possible representation deficiencies that can occur during system design and data production are identified. 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.

A different approach defining and managing data quality is used by [ENG99]. He identifies various categories for data quality, e.g. data definition quality, information architecture quality, data content quality and data presentation quality. For each category a list of detailed quality attributes is set up. The attributes are overlapping and there is no description of the numerous connections and dependencies between these attributes.

[WAN96b] worked out a list of general data quality characteristics based on a multi-staged empirical survey. The main analysis is based on 355 returned questionnaires (rate of return 24%). Accuracy and correctness were identified as the quality attributes most important to end-users. The final results of this study are four categories (Intrinsic, Contextual, Representational, Accessibility), each containing several detailed quality attributes.

[JAR99] propose 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. This results in prototypical goal hierarchies for each of these user roles [JAR00].

In this article an alternative concept for structuring data quality based on an approach by [GAR84] is proposed. According to [GAR84] 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 data quality and will therefore not be considered further. *Product-based* definitions are quite different; they view quality as a precise and measurable variable. Quality is precisely measurable through inherent characteristics of the product. By contrast, *User-based* approaches assume that quality is determined by the user. According to these approaches individual consumers have individual wants, and the product which satisfies them best is assumed to have the highest quality. This is an 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 process. Virtually all manufacturing-based definitions regard 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 (except 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.

Taking this approach, three relevant levels for data quality in data warehouse systems can be identified: The *user-based level* concentrates on the end-user‘s quality demands 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*). Because of different, level specific aims, different measurement methods are adequate to evaluate the particular quality for each 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. Figure 1 shows the connection between the three levels and two factors of data quality. 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 databases 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.

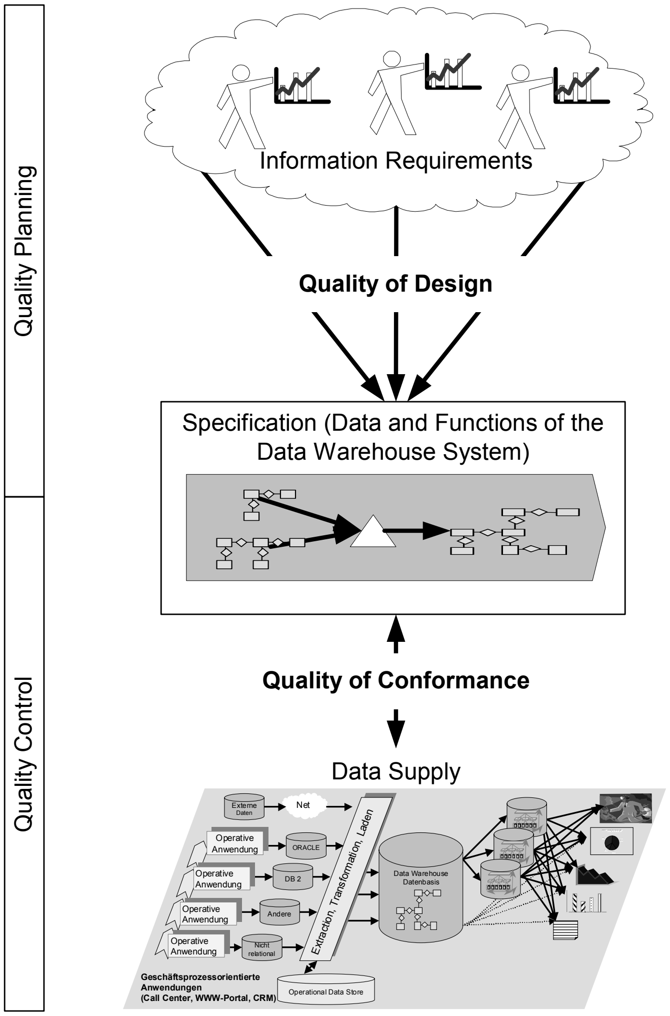

Figure 1: Data quality levels and factors

2 Metadata Based Data Quality System

2.1 Conceptual Framework

The data quality system focuses on the continuous improvement of quality within the scope of a proactive data quality management. The system comprises the whole data warehouse architecture including all applications from operational to analytical information systems. The data quality will be measured and evaluated along the data flow. The metadata management plays a key role in this quality system. Especially metadata about the transformation processes and data schemes are used for gauging data quality. In addition hand-made or tool-supported analysis of data quality experts as well as quality assessments of end-users will be taken into account. Figure 2 shows the architecture and key components of the data quality system for measuring data quality in a data warehouse system.

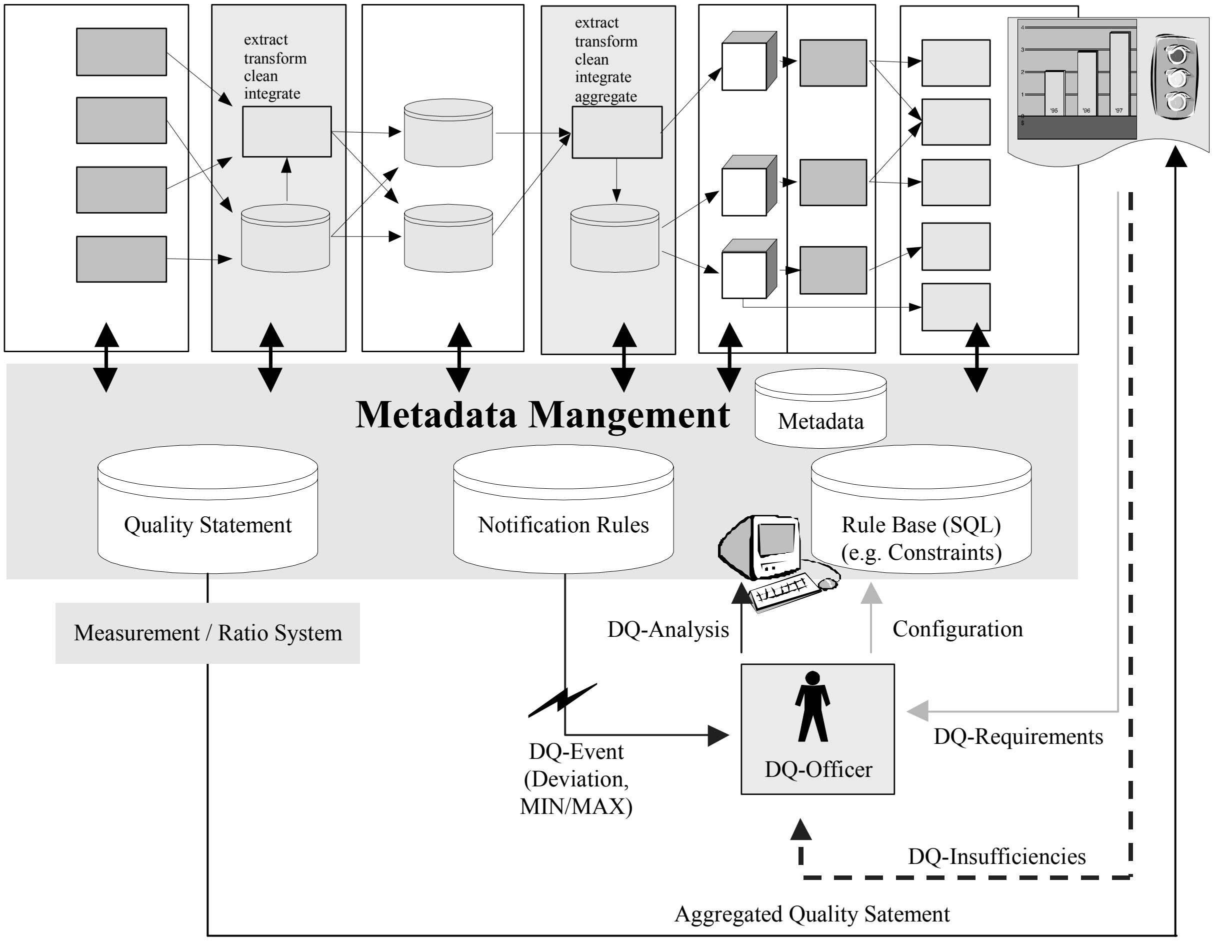

Figure 2: Architecture of the metadata based data quality system

The most important part of the concept is an integrated metadata management component, which contains all important information about data quality. The rule base covers all rules for evaluating the data quality. Not only the conditions and measurement objects but also the time schedules for the execution are specified by the rules. Furthermore notification rules and quality statements are an integral part of the metadata management. If deviations from quality standards or violations of quality rules occur the notification statements define who shall be contacted by which way. For example, an e-mail could be sent to the data quality officer if a special data quality rule is infringed. Then appropriate steps can be taken by the responsible person. The quality statements also contain the results of the data quality measurements and determine in which way they will be presented to the end-users. For example, hierarchical automatic control loops could be used to aggregate low-level quality results to key figures for the users. Based on experience three top level quality statements are sufficient: the data is usable (green), partly usable (yellow) and not usable (red).

2.2 Measuring Data Quality

The goal of quality of design is collecting quality requirements and transforming them into a specification. Data schemes of the data warehouse and functional applications, which define the data objects and their syntax and semantic, are most relevant for the quality of design. Consistency between the schemes has to be ensured. This has a deep impact on the interpretability and usefulness of the data. The quality of conformance tries to assure the compliance of the data with this specification during the operation of the data warehouse. Plausibility, timeliness and availability are the main categories of quality of conformance.

The set of rules for measuring the plausibility of data is based on dependencies and consistency conditions of the real world. They can be used to define semantical constraints for the data. [DAT00; ELM00] distinguish between static, transitional and dynamic constraints. They can be specified for attributes, tuples, relations or the whole database. Additionally, several more types of constraints exist:

- A value is linked to another value e.g., the sum of accounts of a special customer group corresponds to the sum of accounts of another customer group minus the added-up credits of a third customer segment.
- The number of rows of a table depends on the number of rows of another table e.g., the number of accounts is equal or bigger than the number of customers.
- A certain value is time-invariantly e.g., the number of seconds per minute.
- A value varies in the same way as another value over a specific period of time e.g., the number of customers behaves linear to the expenses of the marketing department.

All these types of constraints are based upon the assumption that some properties of the data are time-invariant, in order for them to provide a suitable basis for comparison with other volumes of data. The data used for generating valid conditions is called quality reference data, which is identified and selected by experts with business knowledge. Descriptive statistics and data mining methods are used to derive these conditions and constraints from the quality reference data, which can be applied on other data sets. To obtain universally valid constraints, which can be applied to other volumes of data, slight fluctuations can be flattened by aggregating the data, whereas considerable variations in the data have to be modeled as probability distributions to explicit seasonal or other effects.

The methods of descriptive statistics provide characteristic properties of the data sets, which can be used to analyze and describe the data. The results can be plotted for a better understanding. Data characteristics are then utilized to specify constraints and rules for measuring data quality. [RIC94; MAN01; BOX78] differ between univariate and multivariate techniques. Frequency distributions, empirical variances and standard deviations are limited to only one variable, whereas methods like conditional distributions and regression analysis can include more than one variable. Data mining is used to explore unknown correlations in the data. The methods of inductive learning generate previously unknown links and patterns of large databases. The results can be applied on new data tuples for predictions and classifications. Examples for data mining techniques are cluster analysis, neural networks, association analysis and decision trees [WIT99; MAN01]. They are used to identify segments, classes and association rules in the data, from which quality relevant statements can be derived. Formalized descriptions with specific algorithms for data quality mining can be found in [SOL01; GRI01].

In addition to semantical constraints the data volume is an appropriate item for measuring data quality, especially the timeliness of data. The amount of data handled during the ETL process usually varies within certain limits over a specific period of time and is therefore suitable for gauging data quality. The quantity of data at certain points in time can be measured and the number of records of future ETL processes can be predicted by applying time series analysis. For example, if the data warehouse is only partially loaded the measured data volumes deviate from the predicted value. This indicates that timeliness of data shown to the end-users is impaired. Figure 3 depicts potential methods for measuring data quality particularly the plausibility, timeliness, interpretability and usefulness of the data.

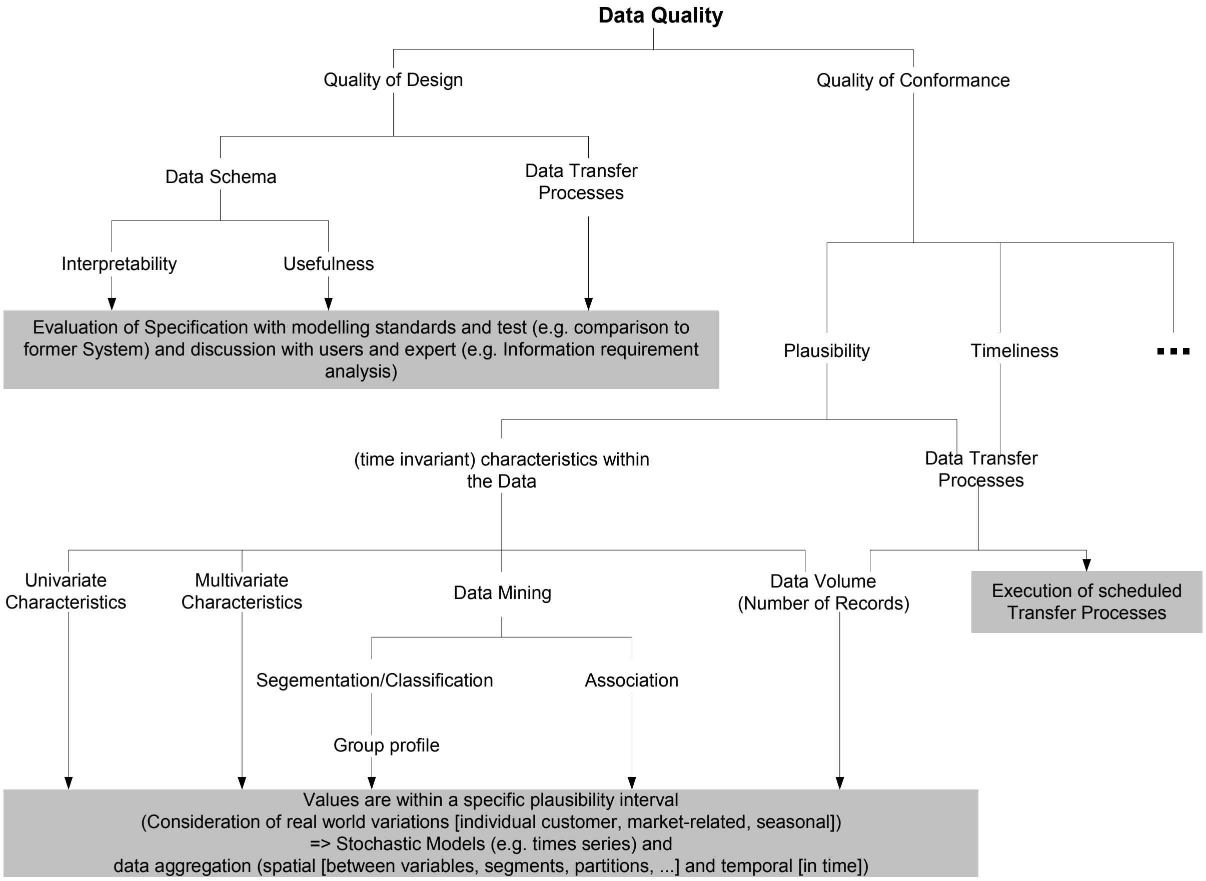
Data Quality
Quality of Design
Quality of Conformance
Data Schema
Data Transfer Processes
Interpretability
Usefulness
Evaluation of Specification with modelling standards and test (e.g. comparison to former System) and discussion with users and expert (e.g. Information requirement analysis)
Plausibility
Timeliness
...
(time invariant) characteristics within the Data
Data Transfer Processes
Univariate Characteristics
Multivariate Characteristics
Data Mining
Data Volume (Number of Records)
Execution of scheduled Transfer Processes
Segementation/Classification
Association
Group profile
Values are within a specific plausibility interval
(Consideration of real world variations [individual customer, market-related, seasonal])
=> Stochastic Models (e.g. times series) and
data aggregation (spatial [between variables, segments, partitions, ...] and temporal [in time])

Figure 3: Measuring data quality

2.3 Practical Experiences

The conceptual architecture of the metadata based data quality system described above is currently implemented in the data warehouse architecture of a large Swiss bank. The project is accompanied and supported by the Competence Center ‘Data Warehousing 2’ and the first implementation release is nearly realized. All components of the proposed framework are part of the first release, only the aggregated view (red, yellow, green) on the quality key figures is not implemented but planned for the next step. Table 1 shows a part of the data quality specification for the bank. Due to privacy concerns the specification is generalized and anonymizated.

Criteria		Approach	Rule
Plausibility	**Domain and attribute**	Value corresponds to a specified data type or format	An order number for a payment transaction consists out of 13 characters. The first 4 characters have to contain a number, all other symbols have to be alphabetic characters.
		Value is within a range of values	The values of the attribute YearOfBirth must be between ‘1890’ and ActualYear.
		Obligatory fields (missing values)	The attribute a of table x must not be NULL or blank.
	Record and relation	Key values are unique	The key attribute of table x must be filled with a unique value.
	Referential integrity	Foreign key relations	The key attribute of table x must exist in table y.
Timeliness		Data volumes (number of rows)	The table x increases by approximately 1000 rows per month. or Between March 2001 and December 2001 table y increased by 10000 to 12000 rows per month. In January 2002 only one-third of the rows of the previous moth are expected.

Table 1: Cutout of the quality specification of a large Swiss bank (generalized)

The verbal rules for measuring data quality are implemented in the rule base as SQL statements. The SQL statements are handmade and especially non-formalized rules stated by business people have to be translated in SQL. For example, the simplified statement for measuring the number of new rows of table x for January 2002 looks like:

```
select count (*)

from   table_x x

where  x.date_per = to_date('31.01.2002','dd.mm.yyyy')
```

Similar SQL statements are possible for all rules of a data quality specification. In contrast to the technical rules business related rules could only be expressed in an iterative and difficult process. Several meetings and internal workshops with developers, business people and analysts were necessary in order to acquire rules with a business context. The evaluation of the data quality of the new data in the staging area is performed on a daily basis. It is the last job before the data is loaded into the data warehouse. The data quality system can process all quality rules as long as they are expressed as SQL statements. The only given limitation is the performance needed for measuring the data quality. Performance problems occurred during the testing phase for very complex rules executed on the data in the staging area. The presentation of previously not known data quality problems to end-users increased their satisfaction as well as their sensibility for data quality aspects. As a consequence of explicate data quality problems, further steps will include measures for improving data quality.

3 Conclusion

The proposed data quality system provides an approach for data quality planning and data quality measuring in data warehousing. The system establishes, as one of the key components for proactive data quality management, the foundation for ensuring high level data quality in data warehouse systems. It provides a way of stating data quality requirements through a rule base (e.g. integrity constrains) and measuring the current quality level by applying these specified rules. The implementation of the system in a large Swiss bank shows the practical capability to measure data quality. The measured quality statements provide quality information to end-users, who do not need detailed knowledge about technical data models and transformation processes. The article shows the basic concept of the components and the rule bases of the data quality system. The practical experiences show that the rules have to be stated by the end-user and may become very complex. Only they are able to state their quality requirements in a non-formalized way and have the business knowledge to understand the data (e.g. semantic constrains). Even if the system is implemented further research is required. For example more detailed research is needed for stating quality rules (e.g. data mining and statistics). Common organizational concepts and methods for data warehouse systems have to be enhanced by data quality aspects (DQ-Officer, information requirements analysis). This approach could be applied to other fields like e-commerce, logistics and knowledge management.

References

[BOX78] G. E. P. Box, J. S. Hunter, W. G. Hunter: Statistics for Experimenters: An Introduction to Design, Data Analysis, and Model Building, John Wiley & Sons, New York et al., 1978.

[DAT00] C. J. Date: An introduction to database systems, Addison-Wesley, Reading, Massachusetts et al., 2000.

[ELM00] R. A. Elmasri, S. B. Navathe: Fundamentals of Database Systems, Addison-Wesley, Reading, Massachusetts et al., 2000.

[ENG99] L. English: Improving Data Warehouse and Business Information Quality. Wiley, New York et al. 1999.

[EPP00] 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, Massachusetts Institute of Technology, Cambridge, MA, 2000, p. 83-96.

[GAR84] D. A. Garvin: What does 'Product Quality' really mean?, in Sloan Management Review, 26(1984), No. 1, p. 25-43.

[GRI01] U. Grimmer, H. Hinrichs: A Methodological Approach to Data Quality Management Supported by Data Mining, in Pierce, E. M., Kaatz-Haas, R. (ed.): Proceedings of the Sixth International Conference on Information Quality, Massachusetts Institute of Technology, Cambridge, MA, 2001, p. 217-232.

[GUT94] T. A. Gutzwiller: Das CC RIM-Refernzmodell für den Entwurf von betrieblichen, transaktionsorientierten Informationssystemen, Physica-Verlag, Heidelberg, 1994.

[HAE98] C. Haeussler: Datenqualitaet, in Martin, W. (ed.): Data Warehousing, ITP GmbH, Bonn, 1998, p. 75-89.

[HEL00] 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.

[HEL01] 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.

[HUA99] K. Huang, Y. Lee, R. Wang: Quality Information and Knowledge, Prentice Hall, Upper Saddle River, NJ, 1999.

[JAR99] 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.

[JAR00] M. Jarke, M. Lenzerini, Y. Vassiliou, P. Vassiliadis: Fundamentals of data warehouses, Springer, Berlin et al., 2000.

[JUR79] 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.

[MAN01] H. Mannila, P. Smyth, D. J. Hand: Principles of Data Mining, MIT Press, Cambridge, MA, 2001.

[RIC94] J. A. Rice: Mathematical Statistics and Data Analysis, Duxbury Press, Belmont, California, 1994.

[SEG96] H. D. Seghezzi: Integriertes Qualitätsmanagement – das St. Galler Konzept, Hanser, Munich and Vienna, 1996.

[SOL01] S. V. Soler, D. Yankelevich: Quality Mining: A Data Mining Method for Data Quality Evaluation, in Pierce, E. M., Kaatz-Haas, R. (ed.): Proceedings of the Sixth International Conference on Information Quality, Massachusetts Institute of Technology, Cambridge, MA, 2001, p. 162-172.

[TAY98] G. K. Tayi, D. Ballou: Examining Data Quality, in Communication of the ACM 41(1998), February, No. 2, p. 54-57.

[WAN96a] Y. Wand, R. Y. Wang: Anchoring Data Quality Dimensions in Ontological Foundations in: Communications of the ACM, 39(1996), No. 11, p. 86-95.

[WAN96b] 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.

[WIT99] I. H. Witten, E. Frank: Data Mining: Practical Machine Learning Tools and Techniques with Java Implementations, Morgan Kaufmann Publishers, San Francisco, 1999.