A Strategy for Managing Data Quality in Data Warehouse Systems

— Work in Progress —

Markus Helfert, Eitel von Maur
Institute of Information Management, University of St. Gallen
Mueller-Friedberg-Strasse 8, St. Gallen, CH-9000, Switzerland
phone: +41-71-224-33 82 fax: +41-71-224-21 89
http://datawarehouse.iwi.unisg.ch
{Markus.Helfert ¦ Eitel.vonMaur}@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 Data Quality Management (DQM) for Data-Warehouse-System will be described. As key part DQM an approach for operative DQM (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. INTRODUCTION

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. (English 1999; Helfert 2000a) In general, ensuring high level data quality is one of the most expensive and time-consuming tasks to perform in data warehousing projects (Mueller 2000; Haeussler 1998). Because of insufficient data quality, data warehouse projects are frequently discontinued (Helfert 2000b).

In several publications (Jarke et al. 2000; Helfert 2000b; Huang et al. 1999; English 1999; Tayi/Ballou 1998) approaches for managing data quality have been suggested, but the question of how to ensure high level data quality in data warehouse systems still remains. A key question in data quality management is the operative tasks of quality planing, specifying processes and measuring them on an integrated and most objective base. An evaluation of major approaches shows that there is still a lack of gathering data quality requirements, quality planing and transforming these requirements into a specification and controlling them. One approach, which was developed at the MIT, models quality requirements based on Entity-Relationship-Models.(Wang et al. 1993) Another approach, which was developed within an European research project, considers technical aspects of integrating quality requirements into meta data management.(Jarke et

al. 2000) Most approaches are based on data quality criteria lists and are not linked to measurement systems with quality indicators. They also lack guidelines and methods for applying them to company specific requirements. However, there is still no adequate model for integrating operative data quality management into data warehouse systems.

2. DATA-WAREHOUSE-SYSTEMS

In today's competitive and global business environment, understanding and managing information is crucial for companies in order for them to make timely decisions and to respond to changing business conditions. Data processing applications have proliferated across a wide variety of systems over the last two decades, complicating the task of locating and integrating data within the enterprise wide information system. They are often developed separately resulting in different data models, data descriptions and interpretations. As a result, to manage and use the data, many organisations today are building data warehouse systems. A data warehouse system supports information processes by creating an integrated database of consistent, enterprise wide and historical data. It integrates data from multiple, incompatible systems into one consolidated database.

Central component of a data warehouse system is the data warehouse data base, which is simply a single, complete, and consistent store of data obtained from a variety of sources. (Devlin 1997) The data base is used by a number of tools and applications which form the data warehouse system.(Winter 2000) Figure 1 gives an overview of the main components of the data warehouse system, which are to be described in the following briefly.

Starting points are the operational systems and external information systems, which act as data suppliers. With the help of a transformation component the data is extracted, transformed and transferred into the central data warehouse data base. Partly, for subject oriented data supply, smaller, redundant views from the enterprise database are stored in so-called data marts. The databases are accessed by the users with a number of end-user tools. These tools reach from creation of reports over ad-hoc-queries up to multi-layered, multidimensional analyses and data mining.

These components form the basis-system of the data warehouse system by providing the data flow from the data source up to the data use. Alongside this, a co-ordination system, generally named meta data management system, exists.(Holthuis 1999) It consists of a meta data base and tools for storage and administrating the system components and data flows.

3. DATA QUALITY IN DATA-WAREHOUSE-SYSTEMS

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 (Bode 1997), quality (Juran 1998) and data quality (respectively information quality) (Huang et al 1999; Tayi/Ballou 1998; Wand/Wang 1996) and therefore it is necessary to clarify these terms. Many approaches do not distinguish between data and information and define data quality and information quality equally. Before defining data quality, in the following a suitable

definition for knowledge, information and data will be introduced and the different views on quality are described.(Bode 1997; Helfert 2001)

3.1. Knowledge, Information and Data

Knowledge is any form of representation of parts of the real or conceptual world in a material media.(Bode 1997) Characteristic of this definition is the representation of real world objects. Knowledge is an image and is not identical with the real world. But however it is related to the real world and thus has some meaning (semantics). Based on this definition, **information** can be understood as a subset of knowledge, which can be expressed in some form of human language.(Bode 1997) Human language is limited to languages for communicating between humans. Following this definition, **data** can be defined as a subset of information, which is oriented to be processed by machines (e. g. applications and data base systems) .(Bode 1997)

3.2. Quality

The term quality is as complex as the term information.(Juran 1998) As a consequence of the discussion on this term, a variety of definition and interpretation approaches exists. The aim of the definition is to reduce the complexity of the quality phenomenon and to attain operational statements.

According to the classification of (Garvin 84) quality approaches can be differentiated into five quality definitions. 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 this thesis. **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 (apart from the transcendent view) are important on different levels of the design process. It is not possible to focus only on one perspective. The different approaches represent the levels of requirement analysis, product and process design and the actual manufacturing process.

3.3. Data Quality

Like the terms above, data and information quality are described in literature through many different views, whereas the user-oriented view dominates. There are many different definitions with a vast quantity of quality criteria.(Wang et al. 2001; Mueller 2000; Naumann/Rolker 2000; Wolf 1999; English 1999; Tayi/Ballou 1998; Jarke/Vassiliou 1997; Wang/Strong 1996; Redman 1996) Result of this work is a multiplicity of criterion lists and classification frameworks for different areas. In the context of the thesis and as basis for developing a data quality model, these approaches are to be examined and classified.

Although the terms data and information quality are used without uniformity, the analysis of the approaches shows a conformance that data quality is determined according to a user-oriented view. This view is concretised through quality criteria, which depend in its meaning and intensity on the application. The approaches do not show any conformance regarding the quality criteria lists, their definitions and systematic. Generally the quality criteria are created intuitively on the basis of experiences (Ballou/Pazer 1985; Laudon1986; Morey 1982), literature (Naumann/Rolker 2000) or by empirical research (Wang/Strong 1996).

Basis of further work is often the quality criteria list from (Wang/Strong 1996). On the basis of these criteria Jarke and Vassiliou suggest a model for quality factors in data warehouse systems.(Jarke/Vassiliou 1997) Main differences to the initial model lie in the greater emphasis on historical as well as aggregated data. Figure 2 illustrates the hierarchy of quality factors used. After describing the term data quality and list and some relevant quality criteria, the concept of data quality management will be described in the next section.

4. DATA QUALITY MANAGEMENT

Quality management includes concepts of quality policy, quality strategy, quality planning, quality control and quality assurance as well as quality improvement. (Juran 1979; Deming 1982; Seghezzi 1996) 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 has already been successfully implemented in the manufacturing sector. Currently the concept of TQM is applied to sectors like the service industry and data quality. (English 1999; Redman 1996; Wolf 1999) Typical for TQM is the orientation on customer requirements, the participation of people, continuos improvement and the comprehensive management approach. All enterprise wide activities are integrated into an enterprise wide structure aiming continuously to improve products, services and process quality and therefore satisfying customer requirements.(Seghezzi 1996) Following the total quality management approach a concept for data quality management for data warehouse systems can be proposed. Three aspects are fundamental to this concept (see also Figure 3) (Helfert/Radon 2000; Wolf99; Huang et al. 1999; English 1999):

- Management has to commit to accept the philosophy of high level data quality and show this commitment in each activity. On the basis of data quality principles and goals a data quality policy and strategy have to be deduced.

- A quality management system with organisational structure, process organisation, standards and specifications, guidelines and rules as its basic elements founding the structure for the

management concept. Frequently inspections ensure continuous improvement of the organisational structure, processes and standards.

- Employees are supported to fulfill the quality processes by adequate methods, techniques and tools.

The operative level of data quality management deals with four main tasks: *Quality planing* gathers requirements and expectations of data consumers, then transfers these requirements into data delivery processes and specifications.(Juran/Gryna 1993; Seghezzi 1996) Therefore quality criteria have to be selected, classified and prioritised.(Wallmueller 1990) *Quality control* controls the data delivery processes and complies the stated specification. Therefore adequate steps have to be identified and implemented. To reach this, product and process quality must be measured and expressed in quantitative indices. Most important techniques of quality control are quality examinations.(Juran/Gryna 1993; Wallmueller 1990) *Quality assurance* aims to detect systematic risks in order to avoid them. *Quality improvement* as the forth task, supports continuos and dynamic quality improvement.

Before analysing and improving data quality it is essential to plan, define and assess quality goals and measure current quality levels (Quality planing and Quality control). Data quality planning and controlling are therefore key success factors of the data quality management concept. This gives the possibility to state current quality levels and compare the results in time. It is possible to identify quality trends and evaluate the effects of quality improvements, which provide the foundation for cost benefit analysis.

4.1. Operative Data Quality Management: Quality planing and measuring

In light of the importance of data quality planning and quality control the thesis will be focused on these two quality management areas. Part of this is a quality model to define quality goals and measure current data quality levels. Therefore, the major goal of the thesis is to develop a suitable data quality model for specification and measurement of data quality in data warehouse systems. This enables data quality planing and data quality control. Figure 4 shows the main structure and focus.

Goal of the quality model is to provide a way to specify quality requirements, create a system for evaluating quality specifications and to measure the resulting data quality. On the basis of quality requirements, which depend on user groups as well as the tasks of the data warehouse system, a framework for a specification have to be developed. A substantial aspect of the data quality model is the decomposition of quality criteria. The general quality term, which is characterised by quality criteria, is decomposed into process and product characteristics. These characteristics are measured by quality indices (defined as quality indicators). Measuring techniques as well as suitable measuring points and times have to be determined.

Starting point for the identification of suitable quality indices is the data warehouse basis-system. As a first step the data delivery processes have to be identified from the data occurrence through the operative system to the data usage. Secondly, appropriate data quality indicators have to be assigned to each data delivery process and data set. Task is to identify typical relations in the data sets and typical characteristics of transformation processes within the basis-system. With the

help of statistics dynamic modifications and inconsistencies in data sets and transformation processes can be detected. The following paragraph shows a simple example of this approach, where research is still to be done.

4.2. A case for the Data Quality Model

To develop a realistic scenario for the proposed data quality model, I worked together with eight large enterprises and we selected an example from an insurance company.

Analytic question: "Number of contracts per region and per sales representative from the perspective of the controlling department. "

This — on a first glance — seemingly simple question turned out to be highly sophisticated and complex to provide through a data warehouse system. First of all, we discussed problems and common data quality requirements, which lead us to data quality measuring approaches.

The main data quality difficulties are different interpretations and multiple applications for this information in different contexts. For example different departments define the term "contract" and the relevant transaction date differently. Problems such as movement of sales representative from one region to another during an accounting period, causes associating difficulties. Through the discussion it turned out that data quality levels are highly dependent on user groups and their intended tasks. Expressed data quality requirements and their related data quality criteria are summarised in the following table (see Table 1):

In a second step we worked out a typical data flow, the data transfer and the transformation processes, which are all shown in Figure 5. First of all, during sales conversations, sales representatives gather contract data and customer information. Regularly this data is synchronised with operative systems (e.g. Mainframe systems). Data typists manually enter non electronic data into the system (e.g. Contracts which could not be entered into the sales representatives' system for some reason). In a further step the final contract is sent to the customer. The customer can accept, change or even cancel the contract (within a given time). Frequently some ETL-Component (Extraction, Transforming and Loading) extracts the new or updated data ("delta data") from the operative systems and transfers it into a staging area. In this temporary data base, quality verifications and improvements are performed. Inconsistencies between new data and data, which is already stored in the central data warehouse data base, could be identified through a link between contract data and sales representatives data already in the data base. A separate data cleansing process handles these quality deficiencies and may possibly improve them. Data which has passed the quality verifications and improvements is then loaded into the data warehouse data base (mostly without any further integrity checks). In a last step, the data is then provided through front-end-tools and subject oriented data marts to data users.

In a next step, we structured the data flow and expressed it in a formal way according to Figure 6, which represents the corresponding model. At the top the relations between the conceptual model, the data model and the physical data store is presented. The left part represents operations on the data store, like insert, update, delete, selection and projection. Usually these operations are performed by SQL and handled by some data base system. Transformations are functions performed on data values, which are elements of data sets. A transformation takes one or more data

sets. The output of transformations are one or more data sets. For example relevant data transformations are gathering data, conversion data values, enrichment and aggregation. (Devlin 1997) One particular transformation is data validation, which takes data values and separates these into two data sets (accepted and not accepted data values). The main control element of the data flow is the session, which runs queries and transformations.

Formalised the data flow in Table 2, we then identified quality indicators (process and product) for measuring data quality. These are shown in Table 3.

5. CONCLUSION

The research so far shows that data quality in data warehouse systems is a crucial issue but also a highly sophisticated task to fulfil. On base of the proposed data quality model it is assumed that there are user group and task dependent quality requirements as well as quality indicators along the data delivery process. These requirements are usually expressed in natural language by end users. To structure and express these user requirements a conceptual modelling language has to be developed and integrated in the conventional data modelling process for data warehouse systems. In further research the quality requirements have to be linked to data quality criteria. These quality indices and their target values have then be linked to the data delivery processes. One more technical aspect is to be integrated into the quality planing and controlling into the conventional meta data management. The data warehouse basis-system could so be controlled and would lead to a controlled and higher data quality in data warehouse systems.

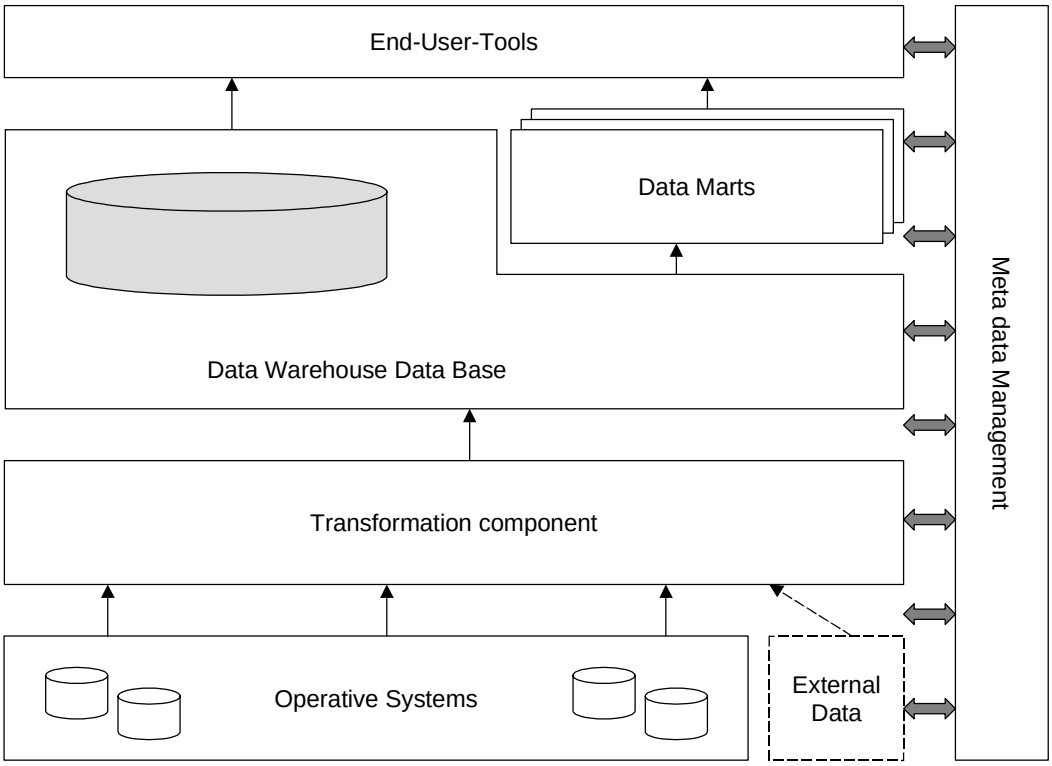

Figure 1: Data Warehouse System (Mueller 2000)

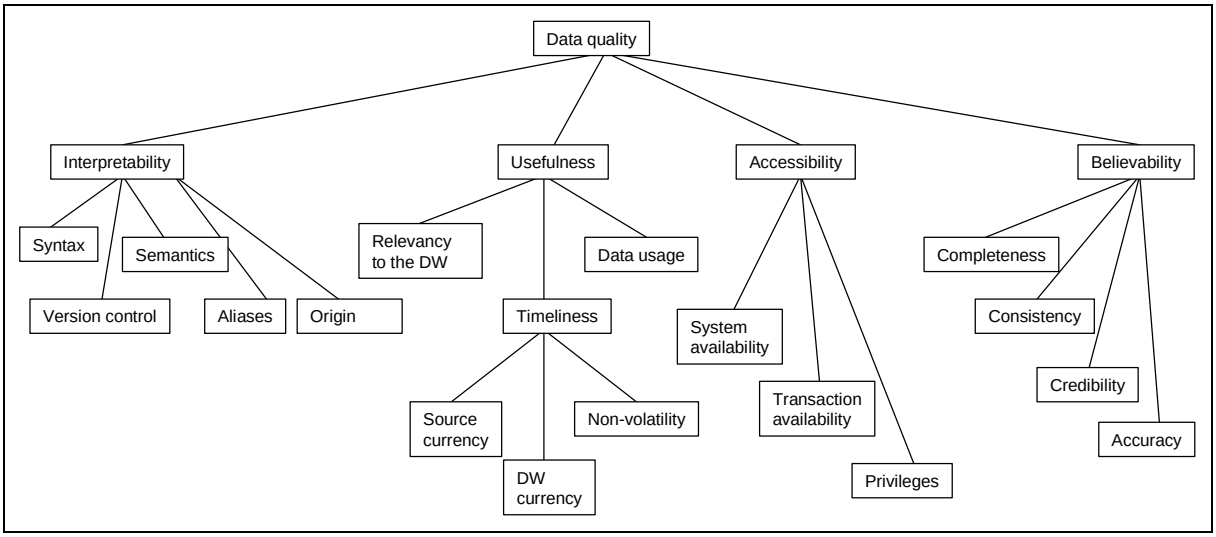

Figure 2: Quality factors for data warehouse systems (Jarke/Vassiliou 1997)

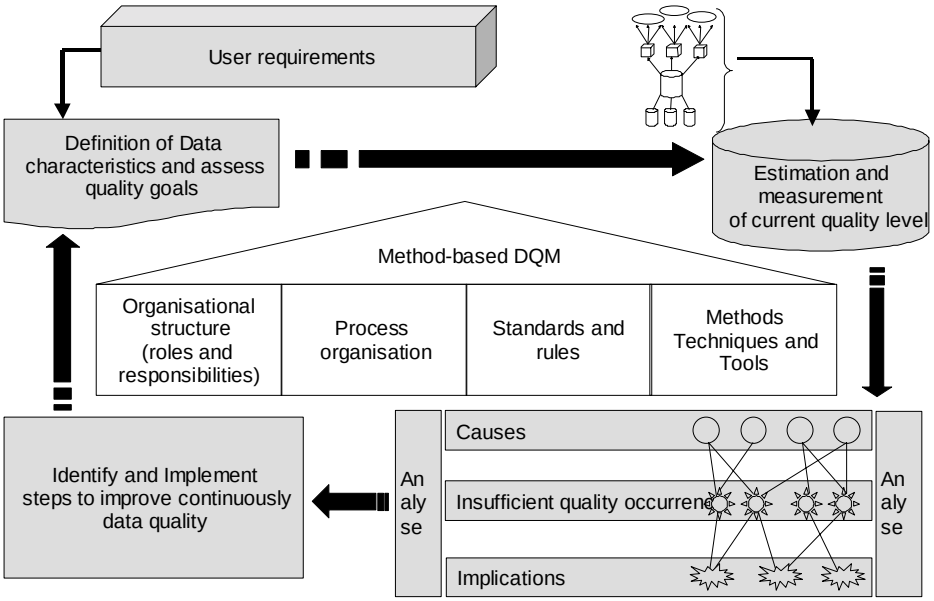

Figure 3: Data Quality Management (Helfert/Radon 2000)

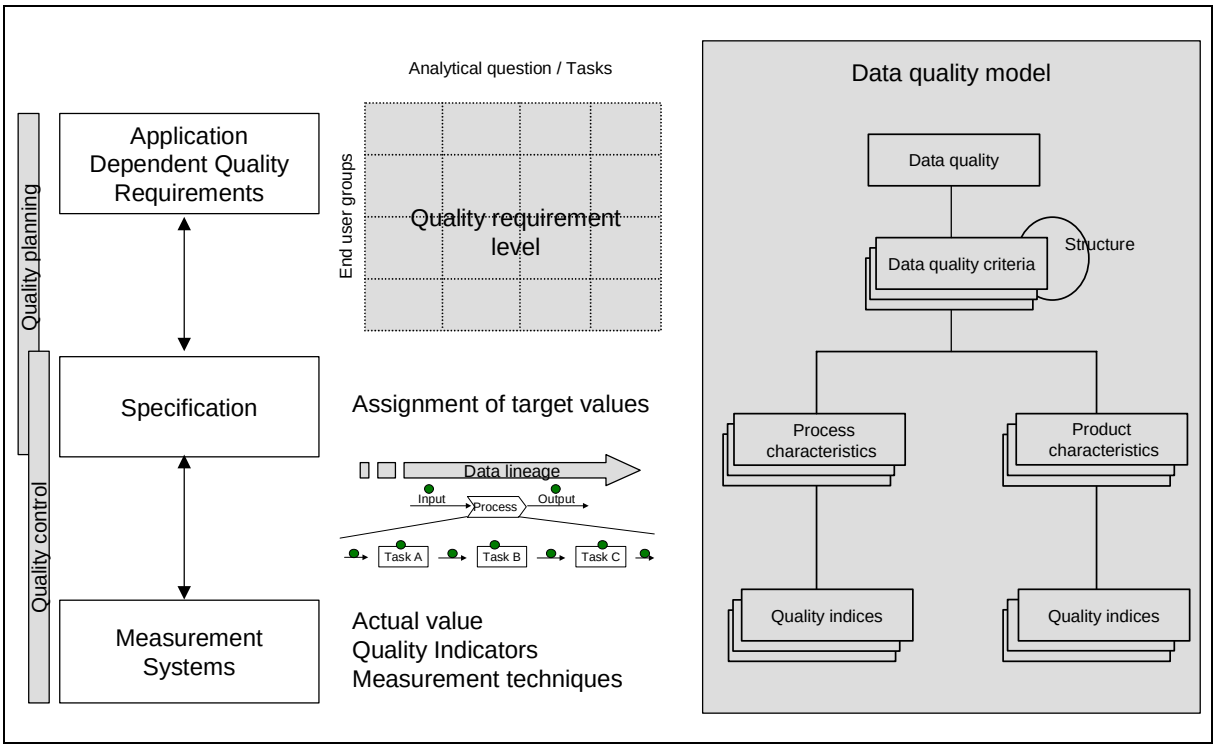

Figure 4: Structure and Focus of the Data Quality Framework

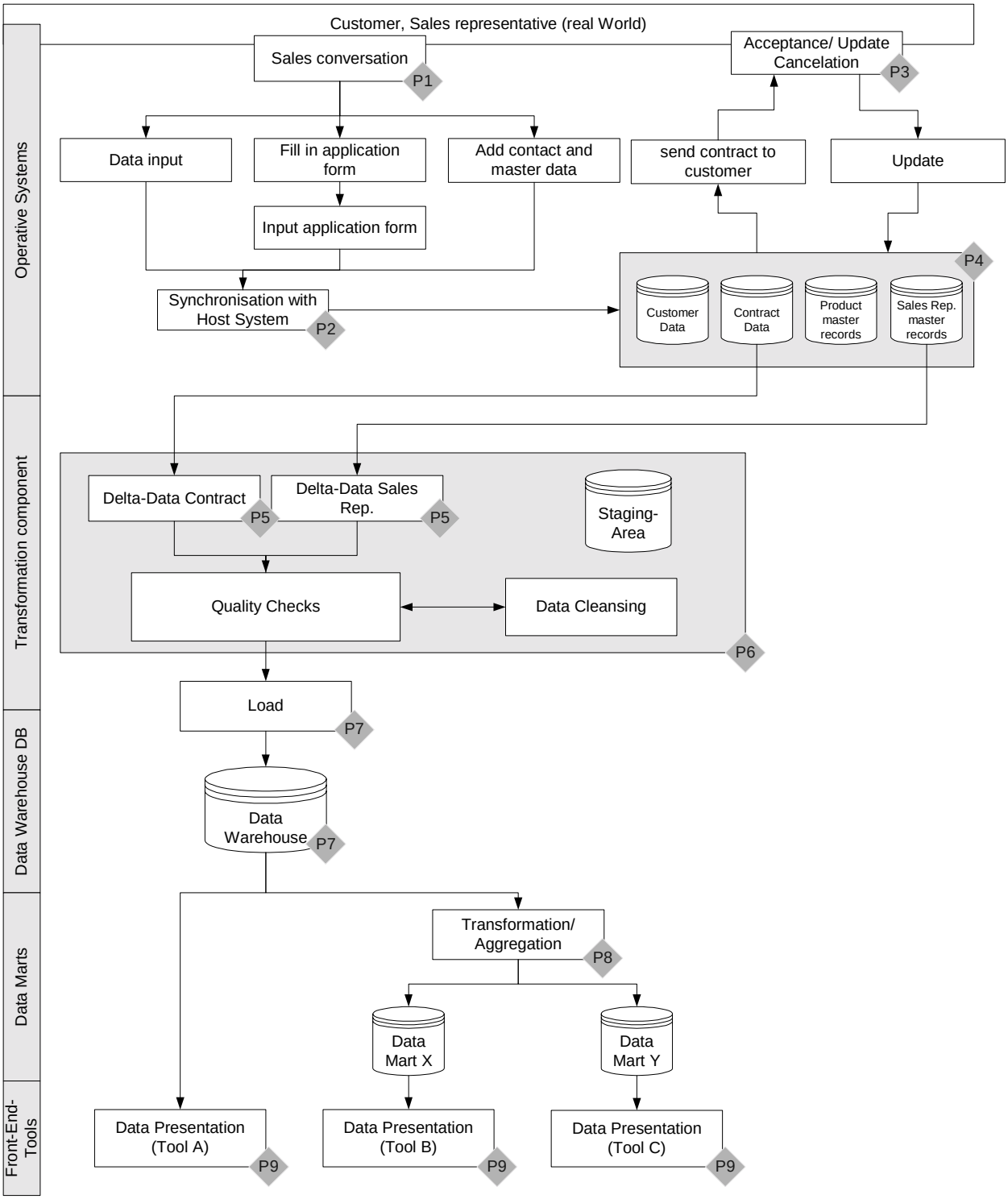

Figure 5: Typical data delivery processes and quality indicators

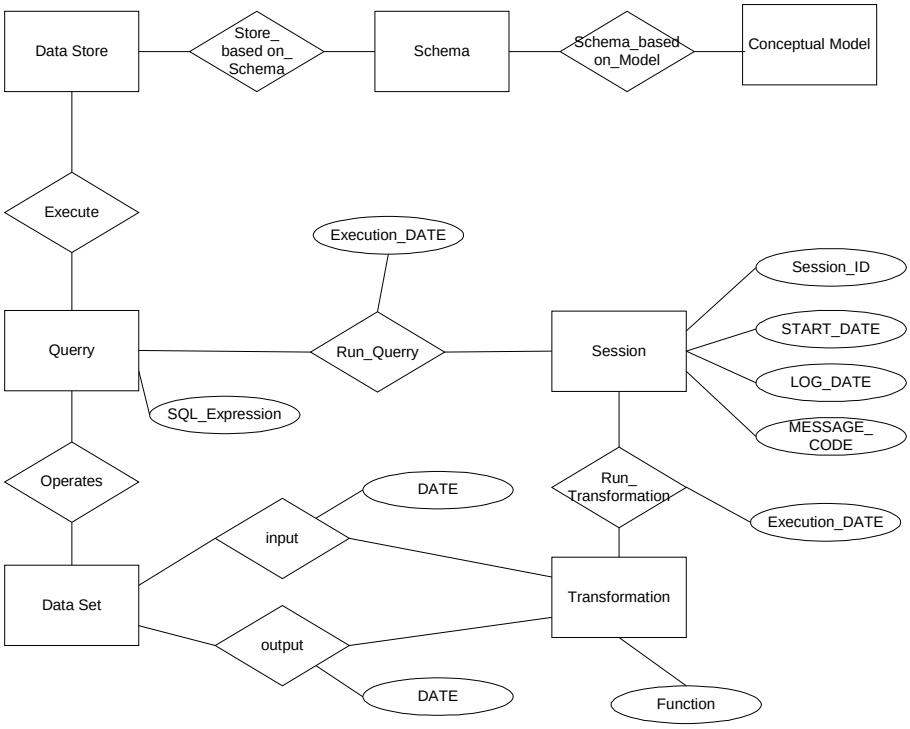

Figure 6: Meta model for Data Flow

Expressed Data Quality Requirements	**Data Quality Criteria**
Maximum allowed divergence to real number of contracts is +/- 2 %.	Accuracy
All sales representatives and regions have to be listed.	Completeness
Format for date has to be “dd/mm/yy”.	Interpretability, Format, Syntax
Information is updated monthly	Timeliness
On the second working day of each month a trend has to be identified (Accuracy +/- 5 of real value).	Accuracy, Timeliness
On the fifth working day of each month the final information has to be provided (In case of later changes reasons have to be given).	Accuracy, Timeliness
Information of the ten best and worst sales representatives have to be accurate.	Accuracy
Number of contracts from new products should be accurate.	Accuracy
Sum of contracts per region and per sales representative have to correspond with the total number of contracts sold.	Accuracy, Consistency
Responds time should be less than three minutes	Timeliness

Table 1: Data Quality Requirements

Data Flow processes	**Labels in Figure 6**
Data Gathering through sales conversation	P1
Synchronisation with operative System	P2
Validation of contract information by customer	P3
Data update and storage in operative Data Bases	P4
Data extraction (delta data)	P5
Data validation and transformation (Cleansing)	P6
Data load and storage in Data Warehouse	P7
Data aggregation and transformation in multi-dimensional Data models	P8
Data presentation	P9

Table 2: Data Flow processes

Data Quality criteria	**Data Quality indicators and measuring points**
Timeliness (currency)	Data gathering date **[P1]** Last execution time / Scheduled time vs. Execution time / Version control through time stamps (protocol evaluation) **[P2, P6, P7,P8]**
Completeness	Completeness of optional data values **[P1, P4, P6]** Numbers of default-values compared to average **[P1, P4, P6]**
Consistency	Plausibility verifications **[P1, P4, P6]**
Accuracy	Frequency of changes **[P4]** Customers' feedback **[P3]** Data user valuation **[P9]**
Interpretability	Data user valuation **[P9]** Domain violation **[P1, P4, P6]**

Table 3: Data quality indicators

6. REFERENCES

(Ballou/Pazer 1985) Ballou, D. P.; Pazer, H. L.: Modeling Data Process Quality in Multi-input, Multi-output Information Systems. In: Management Science 31 (1985) 4, pp. 150-162.

(Bode 1997) Bode, J.: Der Informationsbegriff in der Betriebswirtschaftslehre. In: Schmalenbachs Zeitschrift für betriebswirtschaftliche Forschung, 49 (1997) 5, pp. 449-468.

(Deming 1982) Deming, W. E.: Quality, Productivity and Competitive Position. Cambridge, 1982.

(Devlin 1007) Devlin, B.: Data Warehouse: From architecture to implementation. Addison-Wesley Longman, Reading, MA et al. 1997.

(English 1999) English, L.: Improving Data Warehouse and Business Information Quality. Wiley, New York 1999.

(Garvin 1984) Garvin, D. A.: What does "Product Quality" really mean?. In: Sloan Management Review 26 (1984) 1, pp. 25-43.

(Haeussler 1998) Haeussler, C.: Datenqualitaet. In: Martin, W. (ed.): Data Warehousing. ITP, Bonn 1998, pp. 75-89.

(Helfert 2000a) 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.

(Helfert 2000b) Helfert, M.: Massnahmen und Konzepte zur Sicherung der Datenqualitaet. In: Jung, R.; Winter, R. (ed.): Data Warehousing Strategie – Erfahrungen, Methoden, Visionen – Springer, Berlin et al. 2000.

(Helfert 2001) Helfert, M.: Managing and Measuring Data Quality in Data Warehousing. In: Proceedings of the World Multiconference on Systemics, Cybernetics and Informatics, Orlando, FL 2001, pp. 55-65.

(Helfert/Radon 2000) 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. Massachusetts Institute of Technology, Cambridge, MA 2000, pp. 109-125.

(Holthuis 1999) Holthuis, J.: Der Aufbau von Data Warehouse-Systemen: Konzeption – Datenmodellierung – Vorgehen. Dt. Univ.-Verlag / Gabler, Wiesbaden 1999.

(Huang et al. 1999) Huang, J.; Lee Y. W.; Wang R. Y.: Quality Information and Knowledge. Prentice Hall, Upper Saddle River, NJ 1999.

(Jarke et al. 2000) Jarke, M.; Lenzerini, M.; Vassiliou, Y.; Vassiliadis, P.: Fundamentals of data warehouses. Springer, Berlin et al. 2000.

(Jarke/Vassiliou 1997) Jarke, M.; Vassiliou, Y.: Foundations of Data Warehouse Quality – A Review of the DWQ-Project. In: Strong, D. M., Kahn, B. K. (ed.): Proceedings of the 2nd International Conference on Information Quality, Cambridge, MA 1997, pp. 299-313.

(Juran 1979) Juran, J. M.: Quality Control Handbood. 3rd ed. New York 1979.

(Juran 1998) Juran, J. M.: How to think about Quality. In: Juran, J. M., Godfrey A. B. (ed.): Juran's quality handbook, 5th ed., McGraw-Hill, New York 1998, pp. 2.1-2.18.

(Juran/Gryna 1993) Juran, J. M.; Gryna, F. M.: Quality Planing and analysis: from product development through use, McGraw-Hill, New York 1993.

(Laudon 1986) Laudon, K. C.: Data quality and due process in large interorganizational record systems. In: Communication of the ACM 29 (1986) 1, pp. 4-11.

(Morey 1982) Morey, R. C.: Estimating and improving the quality of information in the MIS. In: Communication of the ACM 25 (1982) 5, pp. 337-342.

(Mueller 2000) Mueller, J.: Transformation operativer Daten zur Nutzung im Data Warehouse. Dt. Univ.-Verlag / Gabler, Wiesbaden 2000.

(Naumann/Rolker 2000) Naumann, F.; Rolker, C.: Assessment Methods for Information Quality Criteria. In: Proceedings of the 2000 Conference on Information Quality, Cambridge, MA 1999, pp. 148-162.

(Redman 1996) Redman, T. C.: Data quality for the information age. Artech House, Norwood 1996.

(Seghezzi 1996) Seghezzi, H. D.: Integriertes Qualitätsmanagement: das St. Galler Konzept. Hanser, Munich et al. 1996.

(Tayi/Ballou 1998) Tayi, G. K.; Ballou, D.: Examining Data Quality. In: Communication of the ACM 41 (1998) 2, pp. 54-57.

(Wallmueller 1990) Wallmueller, E.: Software-Qualitaetssicherung in der Praxis. Hanser, Munich et al. 1990.

(Wand/Wang 1996) Wand, Y.; Wang R.: Anchoring Data Quality Dimensions in Ontological Foundations. In: Communications of the ACM 39 (1996) 11, pp. 86-95.

(Wang et al. 1993) Wang, R. Y.; Kon, H. B.; Madnick, S. E.: Data Quality requirements analysis and modeling. In: Proceedings of the 9th international conference on data engineering (ICDE), IEEE Computer Society, Vienna 1993, pp. 670-677.

(Wang et al. 2001) Wang, R. Y.; Ziad, M.; Lee, Y. W.: Data Quality. Kluwer Academic Publishers, Boston et al. 2001.

(Wang/Strong 1996) Wang, R. Y.; Strong, D. M.: Beyond Accuracy: What Data Quality Means to Data Consumers. In: Journal of Management Information Systems, 12 (1996) 4, pp. 5-33.

(Winter 2000) Winter, R.: Zur Positionierung und Weiterentwicklung des Data Warehousing in der betrieblichen Applikationsarchitektur. In: Jung, R.; Winter, R. (ed.): Data Warehousing Strategie: Erfahrungen, Methoden, Visionen. Springer, Berlin et al. 2000, pp. 127-139.

(Wolf 1999) Wolf, P.: Konzept eines TQM-basierten Regelkreismodells fuer ein „Information Quality Management“ (IQM). Praxiswissen, Dortmund 1999.