An ISO 9001:2000 Compliant Quality Management System for Data Integration in Data Warehouse Systems

Holger Hinrichs
OFFIS
Escherweg 2, 26121 Oldenburg, Germany
holger.hinrichs@offis.de

Thomas Aden
University of Oldenburg
26111 Oldenburg, Germany
thomas.aden@informatik.uni-oldenburg.de

Abstract

In modern information systems, the topic of data quality becomes more and more important due to increasing analysis demands. This holds especially for data warehouse systems and similar systems that provide data integrated from heterogeneous sources. Although a large variety of extraction-transformation-loading tools supporting data integration is available, there is still no process model defining which integration steps should be done in which order to best fulfil the users' needs. Our research project CLIQ is supposed to close this gap. In CLIQ, the integration process is being viewed as a kind of production process. This view enables us to apply concepts of quality management known from the manufacturing/service domain. More precisely, we developed a quality management system for data integration that meets the requirements of the recent ISO 9001:2000 standard. This paper presents our approach and describes its added value compared to traditional approaches to data integration.

1 Introduction

The increasing popularity of data warehouse systems [Inm92] reflects the rising need to make strategic use of data integrated from heterogeneous sources. The integration of heterogeneous data – an integral part of data warehousing – raises the problem of data deficiencies like inconsistent, missing, or obsolete values, unintentional redundancy etc. So if data that do not suffice given user requirements are used for decision support, the so-called "garbage in, garbage out" effect may occur, possibly with serious consequences. To prevent this, the data integration process has to ensure that subsequent analysis tasks work on approved data exclusively. Data that do not fulfil the requirements have to be either improved or singled out. The entire integration process must align with the fulfilment of users' needs. Our research project CLIQ (Data Cleansing with Intelligent Quality Management) is dedicated to this topic.

The scenario we just described can be viewed as a data quality problem. In [ISO00a], the term "quality" is defined as the "degree to which a set of inherent characteristics fulfils requirements". Applied to the context of data warehousing, the subject matter under consideration is "data", more precisely a fragment of a database, below termed "data product". So which characteristics (with regard to the above definition) does such a data product possess? In the literature, several classifications of data characteristics (often called data quality dimensions) can be found [Wan98, JJQV98, NLF99, Sha99]. Unfortunately, there is no commonly accepted classification. In CLIQ, we adopted the classification depicted in Fig. 1, which we consider to be a "best of breeds" concept (see [Hin01]).

Quality characteristics form the backbone of *quality management (QM)*, defined in [ISO00a] as "coordinated activities to direct and control an organization with regard to quality". In accordance with [ISO00a], QM consists of the following activities:

- *Quality policy*
  Establishing overall quality-related intentions and goals of an organization.

Proceedings of the International Workshop on Design and Management of Data Warehouses (DMDW'2001)
Interlaken, Switzerland, June 4, 2001
(D. Theodoratos, J. Hammer, M. Jeusfeld, M. Staudt, eds.)
http://sunsite.informatik.rwth-aachen.de/Publications/CEUR-WS/Vol-39/

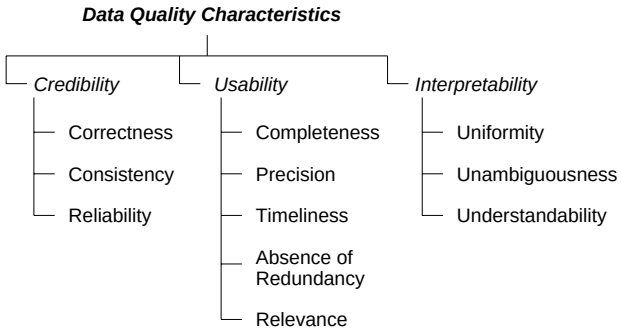

Figure 1: Categories of Data Quality Characteristics

- *Quality planning*
  Setting quality objectives and specifying processes and related resources necessary to fulfil these objectives.

- *Quality control*
  Executing processes to fulfil quality requirements.

- *Quality assurance*
  Providing confidence that quality requirements will be fulfilled.

- *Quality improvement*
  Increasing the ability to fulfil quality requirements.

The system within which QM is done is called *quality management system (QMS)*. As Fig. 2 shows, the data integration process in a data warehouse system can be viewed as some kind of production process, with the heterogeneous source data corresponding to raw materials provided by suppliers (in this case: data sources), the integration process corresponding to the production process, and the consolidated data corresponding to products. As (material) products, data products can be used by customers (in this case: data analysts) whose expectations and needs have to be met.

Following this analogy, we can adapt the well-established QM concepts known from the manufacturing/service domain to the context of data integration, hereafter called *data quality management (DQM)*.

One important aspect of DQM is data quality measurement. As DeMarco stated ”You cannot control what you cannot measure” [DeM82], we need metrics at hand to calculate the degree of conformance with given requirements. In the manufacturing domain, we have to measure characteristics like lengths, weights, speeds, etc. In data warehousing, on the other hand, we have to measure characteristics like consistency, timeliness, and completeness of data products (cf. Fig. 1). Unfortunately, metrics suitable for these characteristics are still a matter of research. Most of the data quality metrics proposed in literature, e. g.

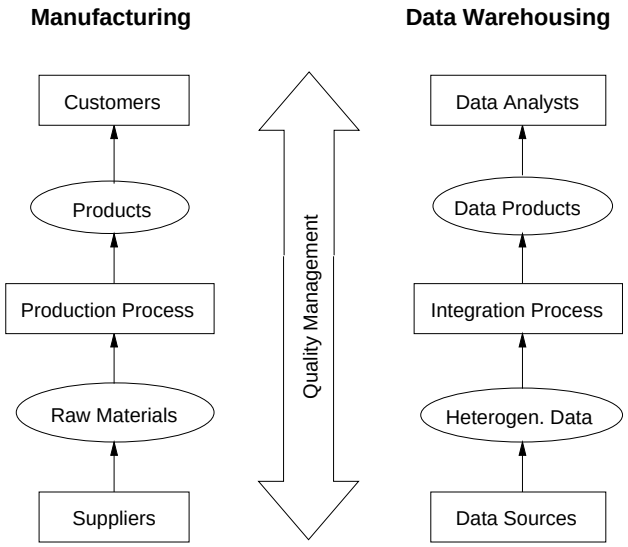

Figure 2: Manufacturing and Data Warehousing

[BT99, NLF99, Sha99], are either trivial or cannot be automated. However, since modern data warehouses store gigabytes up to terabytes of data, a manual quality control is not feasible at all.[1] In [Hin01], we propose a well-defined set of metrics along with automated measuring methods for a subset of the data quality characteristics mentioned in Fig. 1. These are tailored to relational databases as they prevail in the back-end area of current data warehouse systems.

Another major question is how data quality can be improved. Two basic approaches are possible:

- *Symptom-oriented*
  In this case, data cleansing methods are applied to deficient data to improve their quality as far as possible. If data cannot be made conformable with given requirements, they have to be flagged and then either be sorted out or be released for restricted use. In [ISO00b], these aspects are summarized as ”control of nonconforming product”.

- *Cause-oriented*
  In this approach, the business processes which ”produce” the data (especially data acquisition, transformation, and consolidation processes) are tuned continually in such a way that the quality of produced data increases more and more. Causes of nonconformities should be eliminated appropriately (in respect of their effects) in order to prevent recurrence. Furthermore, the QMS itself should be the subject of permanent efforts to improve its effectiveness, efficiency, and/or traceability. In [ISO00b], the entirety of these actions is called ”(quality) improvement”.

[1] A fully automated DQM, however, is also unrealistic. Human interaction will be inevitable when conflicts cannot be solved automatically.

Of course it is always better to strike at the root of a problem, which means in this case to optimize the business processes and/or the QMS. Unfortunately, cause-oriented DQM is often hindered by the fact that the processes in question are outside the optimizer's sphere of influence. Legacy systems e. g., which often deliver low quality data (due to a lack of input controls etc.), usually cannot be extended by appropriate integrity rules belatedly. Besides, many data deficiencies cannot be detected until (resp. are raised by) data integration, e. g. duplicate records from different sources. As a consequence, symptom- and cause-oriented steps should be implemented in combination, complementing each other.

The remainder of this paper is structured as follows: In Sect. 2, an overview of the ISO 9001:2000 standard is given. A QMS for data integration which meets the requirements of this standard is presented in Sect. 3. Subsequent to a discussion of related work in Sect. 4, the paper concludes with a rating of our approach and an overview of future work in Sect. 5.

2 The ISO 9001:2000 Standard

The ISO 9000 family of standards has been developed by the International Organization for Standardization (ISO) to assist organizations in implementing and operating effective quality management systems. First introduced in 1987 and revised in 1994, the current revision was published in December 2000, called ISO 9000:2000. It comprises the following standards:

- *ISO 9000:* Fundamentals and vocabulary
- *ISO 9001:* Requirements
- *ISO 9004:* Guidelines for performance improvements
- *ISO 19011:* Guidelines on quality and environmental management systems auditing.

Some excerpts of ISO 9000:2000 have already been cited in the previous section.

In this section, we give an introduction to ISO 9001:2000 (the postfix ":2000" will be omitted below), since it is the standard most relevant to CLIQ. ISO 9004 and 19011 are beyond the scope of this paper.

ISO 9001 promotes the adoption of a process approach when developing, implementing, and improving a QMS to enhance customer satisfaction by meeting customer requirements [ISO00b]. An organization has to identify various activities, link them together and assign resources to them, thus building up a system of communicating processes. Figure 3 illustrates such a process-based QMS. Customers play an important role in this model since their requirements are used as input to the product realization process and their satisfaction is continually analyzed.

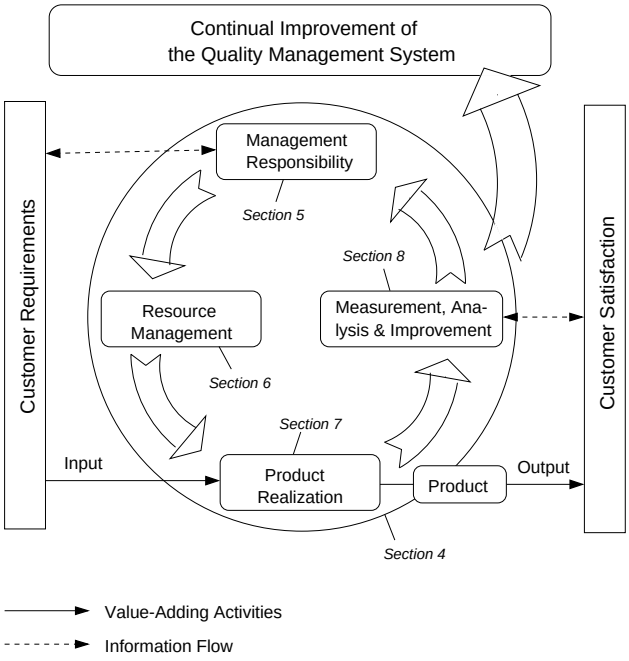

Figure 3: Model of a Process-Based QMS [ISO00b]

ISO 9001 is made up of eight sections, the first three of which contain general information about the scope of the standard, normative references and terms. Sections 4 to 8 describe requirements for a QMS and its components, as indicated in Fig. 3. The structure of these sections is shown in Fig. 4. We will refer to the single sections when we discuss the ISO 9001 compliance of our QMS in the subsequent section.

3 A Quality Management System for Data Integration

In this section, we present a process model for data integration that defines which integration steps have to be executed in which order to achieve optimal data quality. This process model is enriched with DQM steps to ensure that customer requirements are fulfilled. Integration steps plus DQM steps along with organizational DQM activities form a QMS for data integration, called *data quality management system (DQMS)*. In the context of data warehousing, such a DQMS has to be located at the extraction-transformation-loading (ETL) stage (see Fig. 5), with the data warehouse as target database.

The following two subsections describe the main concepts of our DQMS and its implementation as a software system.

3.1 DQMS Concepts

As Fig. 6 shows, the DQMS should be viewed as an integral part of an organization. It is tightly coupled to the organization's management, its human and technical resources,

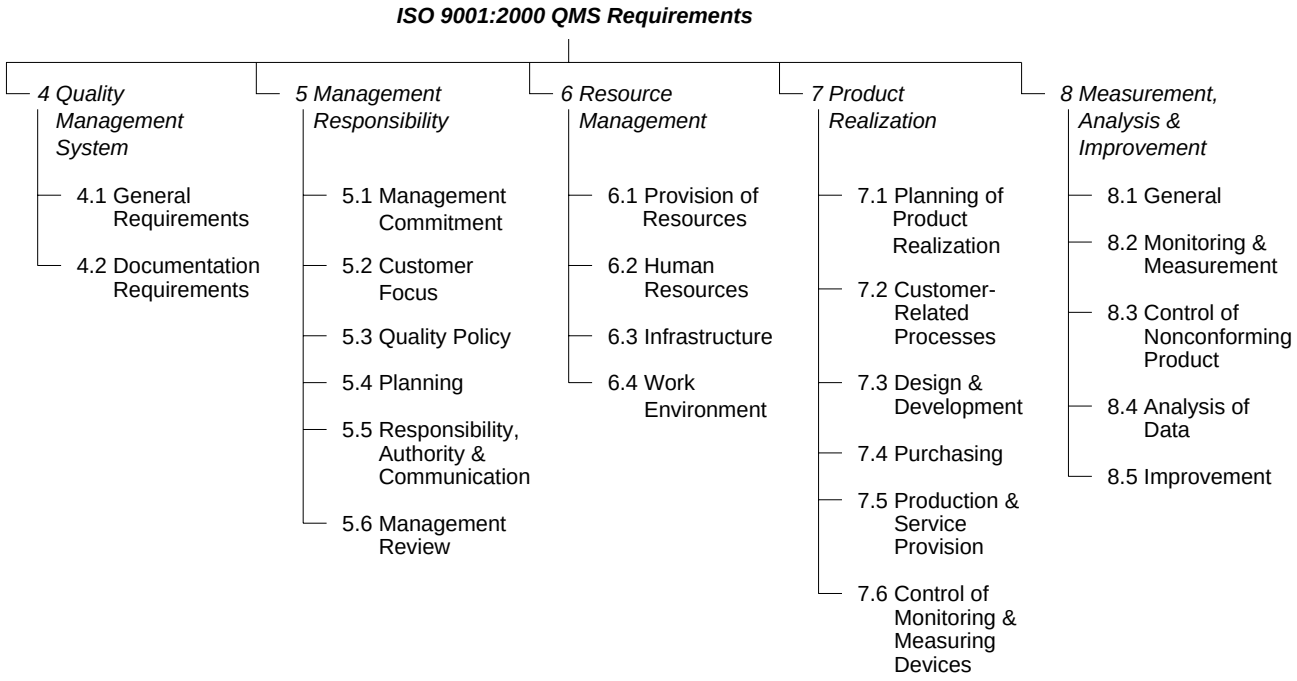

Figure 4: Structure of ISO9001:2000 Sections (Third Level Not Shown)

and – of course – its (data) suppliers and (data) customers. Although the latter two are modelled as external entities in Fig. 6 (to keep the ISO 9001 view), they could also be viewed as internal ones in the data warehousing context (as indicated by the dashed lines): an organization's operational database would be an internal supplier, an organization's data analyst an internal customer. Customers specify quality-related requirements and give feedback concerning their satisfaction with the delivered product. An analogous relationship exists between the organization and its suppliers.

The requirements of ISO 9001 sections 4, 5, and 6 (see Fig. 4) cannot be achieved by the data integration process itself. Instead, they must be met by the surrounding environment as follows:

- The fulfillment of the general requirements of ISO section 4.1 (identification of processes and their interaction, determination of control criteria and methods etc.) has been a basic presupposition of building the DQMS. The same holds for ISO sections 7.1 (planning of product realization) and 8.1 (general measurement, analysis, and improvement requirements).

- We assume that the organization has established and maintained documented statements of the quality policy and quality objectives as well as documentation of procedures used within the DQMS (plus their configurations, where applicable). As in every other ISO 9001 compliant QMS, a quality manual must exist. We further assume that all these documents are stored in a central document management system (electronically), so that – provided that access rights have been granted appropriately – authorized stakeholders may always access the most current (meta) information. Additionally, the document management system is supposed to support version control to meet the requirements of ISO section 4.2.

- Any recordings resulting from integration steps (especially measurement results) are stored in a central metadata repository to meet the requirement to keep recordings "legible, readily identifiable, and retrievable" (ISO section 4.2).

- We assume that the top management of an organization takes steps to meet the requirements of ISO section 5 (management responsibility).

- We assume that the human resources, technical infrastructure, and work environment necessary to operate the DQM processes are provided by the organization as claimed by ISO section 6. Low level resources like CPU time slots, memory allocation, and database access are assumed to be provided by the underlying operating system and DBMS.

In the following, we take a closer look at the integration process itself with regard to its compliance to the remaining ISO 9001 sections 7 (product realization) and 8 (measurement, analysis, and improvement). Since we assume that the organization does not design or develop new (data)

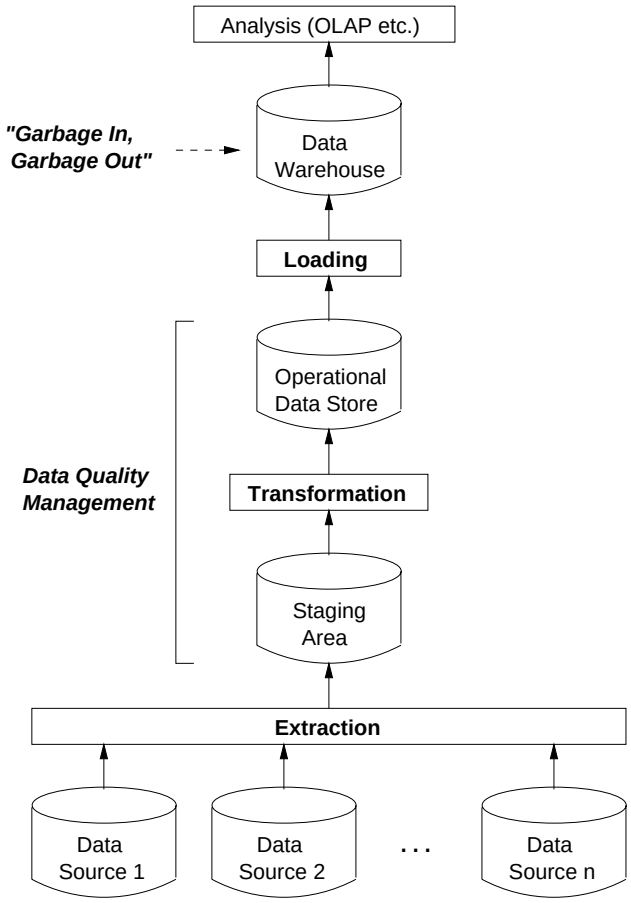

Figure 5: DQM within the ETL Model

products in terms of ISO 9001, we can exclude ISO section 7.3 (design and development) from our discussion.

Step I: Assessment and Review of Requirements

In this step, requirements are determined and maintained (cf. Fig. 6). These include customer requirements in terms of desired values of data quality characteristics, additional implicit requirements as well as statutory and regulatory requirements (e. g. protection of data privacy). The organization has to ensure that it has the ability to meet the specified requirements. The communication between organization and customer is supposed to be handled electronically by means of window dialogs, emails etc. (ISO section 7.2).

Step II: Purchasing Raw Data

According to the requirements specified in the previous step, the organization has to evaluate and select data suppliers. First, a so-called *purchasing information document* has to be set up which specifies requirements the raw data have to meet (cf. Fig. 6), especially concerning data quality characteristics like relevance and timeliness, but also supplier-oriented characteristics like reputation, charges, data acquisition methods, QMS features etc. This document must then be sent to potential suppliers. Based on an evaluation of the suppliers' responses, suppliers have to be selected, supposed their products pass appropriate inspections (ISO section 7.4).

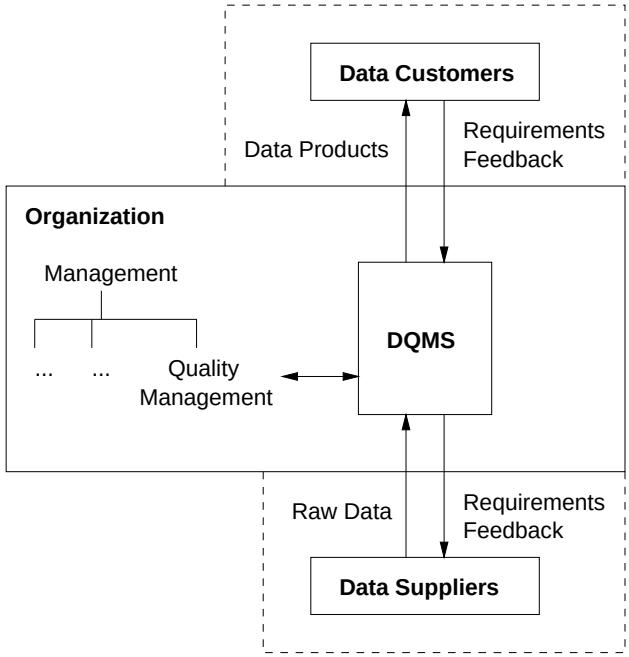

Figure 6: Embedding a DQMS into an Organization

Compared to the purchasing process in manufacturing, there are some peculiarities in the data warehousing context: in many cases, there will be just one data supplier offering a certain contribution to the warehouse data (e. g. the operational databases of the organization itself), i. e. the organization cannot choose between alternative suppliers. Furthermore, in contrast to raw materials in manufacturing, raw data from one supplier may show significant deficiencies in the first place (e. g. missing values), but when being combined with (complementary) raw data from another supplier, these deficiencies may disappear. The organization has to be aware of these special characteristics.

Steps I and II rather belong to the *modelling* phase than to the *operational* phase of the data warehouse system. Therefore, they are arranged *orthogonally* to the operational steps described below.

Assuming that steps I and II have been completed successfully, the data warehouse system may start operations. In CLIQ, we concentrate on the transformation phase of the ETL model, assuming that extraction and loading functionality is provided by third party tools (see also Sect. 4). Each integration process which is to be executed within the data warehouse system should pursue the following phase model (see Fig. 7) in order to meet the ISO 9001 requirements.

Step 1: Unification of Representation

We assume that previously extracted source data are being stored in so-called *source data buffers* (SDB, one per source). SDBs correspond to source-specific "raw material stores", from which heterogeneous data can be fetched and then processed. Each SDB implements a relational schema and thus hides the data model of the respective source (flat file, relational, object-oriented, etc.), simplifying the subsequent integration steps.

In this step, source data are moved from their SDBs into a temporary storage called *transformation area* (also called *staging area* [Kim98]). The transformation area is assumed to have a global relational schema that is covered (in terms of content) by the local SDB schemata. The main task of this step is to map the heterogeneous source data onto the uniform data structures of the transformation area. The following mapping tasks are especially important:

- Generating global IDs and linking them to local keys
- Unifying character strings syntactically (with regard to umlauts, white spaces, etc.)
- Unifying character strings semantically (in case of synonyms)
- Decomposing complex attribute values into atomic ones (e. g. addresses)
- Aggregating atomic attribute values into complex ones (e. g. date values)
- Converting codings (e. g. gender values m/f to 1/2)
- Converting scales (e. g. inch to cm).

To specify the required transformations, conventional ETL tools (cf. Sect. 4) can be used. After executing a transformation, data are written to the transformation area and deleted from their respective SDB.

After this step, the newly extracted records can be identified via their global IDs. Traceability (according to ISO section 7.5) is ensured by a look-up table linking global IDs to local keys (plus their data source ID).

Step 2: Statistical Process Control

After unification, a *statistical process control (SPC)* is done on the transformation area data, based on classical SPC theory [She31]. The idea is to compute attribute-specific statistical key figures (mean, variance, etc.) out of the transformation area data and log them over time. The newly computed key figures are then compared to the formerly logged key figures (used as "expected values"). By doing so, cardinal data deficiencies (e. g. transfer errors) can be detected at a very early stage. If required, i. e. if a priori defined attribute-specific limits are exceeded, an interceding action has to be executed in order to reestablish statistical control, e. g. deleting data from the transformation area and initiating a new data transfer after the problem has been solved.

Step 3: Domain-Specific Consistency Checking

In this step, the transformation area records are being checked with regard to consistency by means of domain-specific knowledge. The latter should be represented in such a way as it can be processed automatically. Different types of representation are possible, especially:

- *Rules*, e. g. "IF DeliveryDate $<$ PurchaseOrderDate THEN Error (SeverityCode 0.7)"
- *Look-up tables*, e. g. bank code registers
- *Regular expressions*, e. g. for phone numbers or article codes
- Arbitrary domain-specific functions.

In case of a detected inconsistency, an appropriate action has to be executed, e. g. by generating an error message or warning. However, the checking process – usually done as a batch run – should not be aborted.

Step 4: Inquiries and Postprocessing

Provided that appropriate domain knowledge is available, the major proportion of inconsistencies can be *detected* automatically, as described above. However, very few inconsistencies can be *corrected* automatically. Consequently, if an inconsistency is not tolerable, an inquiry has to be sent to the affected data source (generated automatically out of an error message, if possible). Depending on the business process implemented, the source may send corrected records to its SDB (then continuing with step 1) or directly to the transformation area. In the latter case, a (preferably automated) postprocessing of corrected records has to be done in the transformation area, followed iteratively by step 3.

Step 5: Record Linkage

The goal of this step is to detect duplicate records, i. e. records that describe the same real world entity, both within the transformation area and between the transformation area and the so-called *target area* (also called *operational data store* [Kim98]) in which the consolidated data will be stored in the end.

Two types of record linkage processes should be distinguished: If a record in the transformation area is just an incremental update of another record already stored in the target area, the linkage can be done via their common (global) keys. Due to the heterogeneity of data sources, this does not work when the extensions of data sources overlap

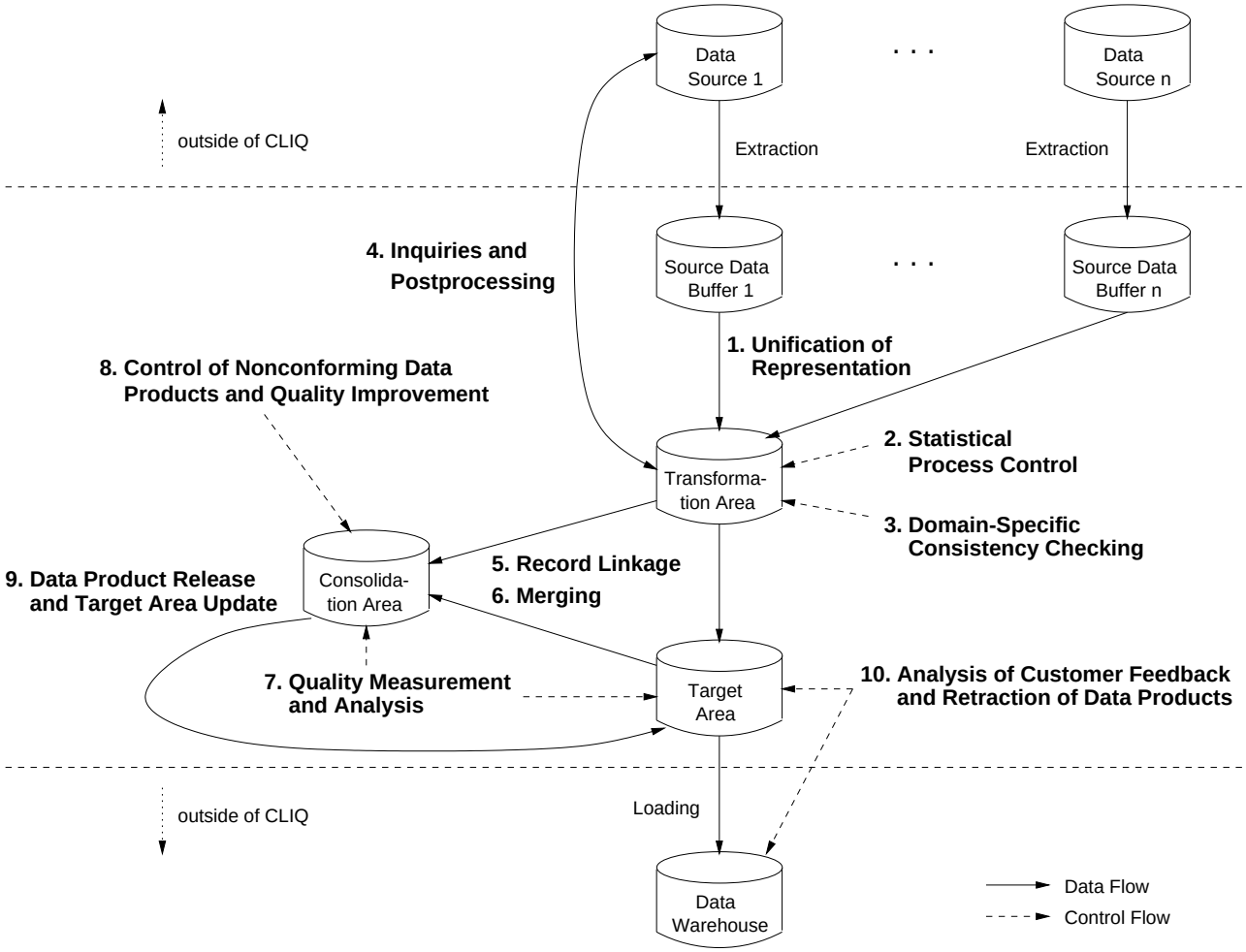

Figure 7: Operational Steps of Data Integration in CLIQ

and, consequently, several source records (with different keys) have to be mapped on the same target area record. The same holds for the case when there are previously undetected duplicates *within* one data source. In these cases, the linkage has to be based on other attributes like name, city, gender, delivery date, etc. To automate this task, several methods have been proposed in literature. The most prominent ones are:

- Probabilistic Record Linkage [Jar89]
- Basic resp. Duplicate Elimination Sorted-Neighborhood Method [HS95]
- Neighborhood Hash Joins [GFSS00].

All these methods result in a set of record pairs potentially belonging together. These pairs, more precisely their transitive closures, now have to be analyzed with respect to whether the linkage is correct or not. Especially in marginal cases, an interactive review will always be inevitable.

Step 6: Merging

Records whose correlations have been validated must now be merged to one single record in order to avoid unintentional redundancy within the data warehouse. Applying certain criteria (information content, attribute-specific priorities on data sources, timeliness etc.), the best pieces of information have to be extracted from the involved records and written into a target record. In our process model, this target record is *not* written to the target area (as one could expect), but into another temporary storage, the so-called *consolidation area*. The target area will be updated not until the very last step of the process model, thus ensuring that only data that have passed all conformance tests (some of which will follow in the subsequent steps) will be written to the target area and thus be made available for analysis tasks.

Those records that merged in a consolidation area record are then deleted from the transformation area. The remaining transformation area records are moved to the consolidation area without any modification. The consolidation area

will now serve as the starting point of the following DQM activities.

Step 7: Quality Measurement and Analysis

In this step, it must be checked if the data in the target area will meet the specified customer (user) requirements (in compliance with ISO section 8.2) *supposed that* it is updated with the current consolidation area data. To do this, the actual quality of data must be measured, using appropriate metrics and measuring software according to ISO section 7.6 (control of monitoring and measuring devices). [2] These measurements[3] must span both the (previous) target area data and the consolidation area data (the target area has *not* been updated yet!).

In a subsequent analysis phase, the results of quality measurements have to be compared to a priori specified quality requirements (resulting from step I). If data do not meet a given requirement, an appropriate action has to be taken (see step 8). Conflicts due to contradicting requirements (e. g. high timeliness vs. high consistency) have to be resolved, e. g. by data replication and different treatment of the replicas.

Furthermore, ISO section 8.2 postulates the measurement of the effectiveness of integration and (data quality) measuring processes.[4] In our context, this can be done by means of *reference data*. For these, both the aspired integration results (characteristics of the resulting data product) and the included quality deficiencies must be known.

Measurement results concerning the effectiveness of products and processes have to be recorded and analyzed (ISO section 8.4), resulting – if necessary – in process improving activities as described below.

Step 8: Control of Nonconforming Data Products and Quality Improvement

In this final step before the target area update, data products that do not conform to given requirements must be treated appropriately, in accordance with ISO section 8.3. The following options may be taken:

- Sort out and re-request data
- Restrict the use of data to specific analysis scenarios, e. g. by flagging
- Eliminate detected nonconformities[5] and then continue at step 7 (subject to ISO section 8.3).

[2] We propose to use the set of metrics and measuring methods developed in CLIQ (see [Hin01]).

[3] Including consistency checks as in step 3 (and, if required, postprocessing as in step 4), since a merging of records may introduce new combinations of attribute values and thus new inconsistencies.

[4] Strictly speaking, these process-oriented measurements should be assigned to a dedicated phase, running parallel to the integration steps.

[5] In CLIQ, we developed a method that uses data mining techniques to eliminate inconsistencies and complete previously missing values semi-automatically (see [HW00] for details).

While all these activities merely tackle the symptoms of a problem, further (cause-oriented) measures may be taken to increase the ability to fulfil quality requirements in the future according to ISO section 8.5:

- Improve the integration process, especially by tuning process parameters (e. g. attribute mappings, SPC parameters, consistency rules, record linkage parameters, merging criteria).
- Improve quality planning and quality control processes, e. g. by finding better means to capture user requirements, by optimizing measurement methods, or by improving feedback methods.

Step 9: Data Product Release and Target Area Update

Depending on the analysis results of step 7, the approved proportion of data is now released (ISO section 8.2), i. e. the affected consolidation area records are flagged as ”passed”. The passed proportion of data is then moved from the consolidation area to the target area, replacing obsolete target area data, if necessary. With this step, the newly integrated data have been made available for customer use. Given a typical data warehouse system, they may now be loaded into the analysis-oriented (e. g. multidimensional) data structures of the data warehouse and attached data marts.

Step 10: Analysis of Customer Feedback and Retraction of Data Products

The organization has to record customer feedback and evaluate it as a measure of the DQMS performance (ISO section 8.2). If a deficiency of a released data product is detected during current use (i. e. by a customer), and this deficiency impairs the usability of the data product significantly, the organization has to retract the product from the target area (and the data warehouse) and ”repair” it if possible (see step 8) before re-releasing it. If necessary, cause-oriented measures should be taken into account (see step 8).

3.2 DQMS Implementation

The CLIQ DQMS has been implemented as a software workbench that offers a continuously quality controlled, extensively automated support of the data integration process. It employs both commercial and self-developed components. Figure 8 illustrates the different layers of the DQMS architecture.

Layer 1 contains the data storage modules (based on Microsoft SQL Server) for business and meta data. Layer 2 offers interfaces to these storage modules. Business data are accessed by way of ODBC, metadata – represented in accordance with the Open Information Model (OIM)

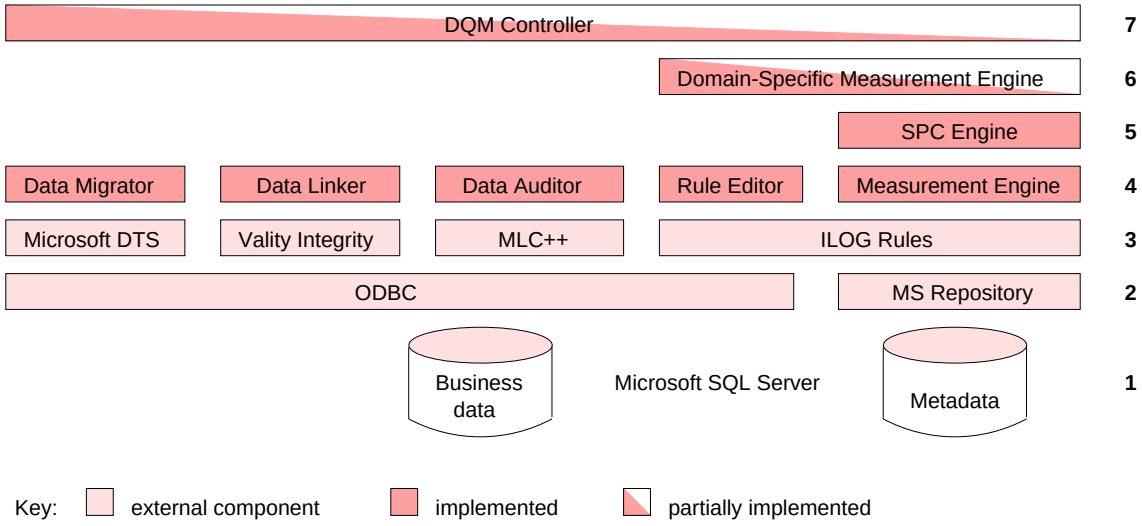

Figure 8: Software Implementation of the DQMS

[MDC01] – by way of the COM-based Microsoft Repository [Ber97].

Layers 4 and 5 form the DQMS core, their submodules being tailored to single steps of the data integration process (Data Migrator: step 1, Data Linker: steps 5 and 6, Data Auditor: step 8, Rule Editor/Measurement Engine: steps 3 and 7, SPC Engine: step 2). On their part, they make use of the commercial components of layer 3, including the Microsoft Data Transformation Services [AL99], the ETL tool Integrity [Val01], the data mining class library MLC++ [KSD96], and the rule engine ILOG Rules [Ilo01].

To be able to integrate domain-specific quality control methods, a so-called Domain-Specific Measurement Engine was introduced in layer 6. Data quality planning, control, and improvement are being coordinated by the DQM controller in the top layer of the software system.

Those software components that have been developed within CLIQ have been implemented using Microsoft Visual C++ 6.0.

4 Related Work

The importance of data quality for organizational success has been underestimated for a long time. For this reason, quality management in information processing is not nearly as established as in other disciplines, e. g. manufacturing. Contemplating the related work concerning DQM, a clear distinction can be made between commercial approaches on the one hand and research activities on the other.

4.1 Commercial Approaches

In recent years, a large number of ETL tools (cf. Sect. 3) hit the market, claiming to simplify the process of populating a data warehouse significantly. But although some of these tools provide sophisticated graphical interfaces and comprehensive libraries of transformation functions, they are insufficient for DQM for the following reasons:

- Their functionality is usually limited to the quality characteristic of uniformity (cf. Fig. 1) by providing typical data migration functions, like mapping of attributes, standardization of addresses, and value conversions.
- They do not support any DQM functions, like quality planning, control, assurance, and improvement.

Nevertheless, an ETL tool can be particularly useful as a single module of an overall DQMS (cf. step 1 in Sect. 3.1). Some prominent examples of commercial ETL tools are:

- DataStage Suite (Ardent Software)
- DQ Plus (Group 1 Software)
- Genio Suite (Hummingbird)
- i. D. Centric Data Quality Suite (Firstlogic)
- InfoRefiner/InfoTransport/InfoPump (Computer Associates, formerly Platinum)
- Integrity (Vality)
- Pure*Integrate (Oracle)
- Trillium Software System (Trillium)
- Warehouse Workbench (Systemfabrik).

A quite extensive list of ETL tools can be found in [Eng99].

4.2 Research Approaches

In 1992, DeLone und McLean [DM92] set up a model of information systems success which included the quality of data as a critical success factor, along with system quality.

In the following years, two major research projects emerged, namely MIT's *Total Data Quality Management (TDQM)* program [Wan98], active since 1992, and the ESPRIT project *Foundations of Data Warehouse Quality (DWQ)* [JJQV98], running from 1996 to 1999.

TDQM claims to establish a theoretical foundation of data quality. Based on the enterprise philosophy of Total Quality Management [Dem86], a so-called TDQM cycle is derived from Deming's classical PDCA cyle (plan, do, check, act) comprising phases of data quality planning, measurement, analysis and improvement. Later, this concept has been extended by [Eng99] and [Hel00]. Additionally, TDQM identifies a set of data quality characteristics (called quality dimensions) based on empirical studies and proposes some simple metrics for data quality measurement. Finally, TDQM provides an approach to enrich the relational data model with data quality information by introducing meta relations. Although TDQM offers some interesting ideas, most of the concepts remain on a quite superficial level, orientating much more to business economics than to information technology.

The DWQ project, on the other hand, is tailored to the data warehousing world. Not only aspects of data quality are considered, but also aspects of schema and software quality. There is a clear distinction between conceptual, logical, and physical data models, the semantics of which are explicitly described as metadata. Integration of source schemata is supported by so-called interschema knowledge networks which specify relationships between schemata. Quality assessment is done using the goal-question-metric (GQM) approach by [BW84]. With GQM, quality goals are specified and then mapped to "quality questions" (in this case: queries on a meta database). To get a response to a quality question, (simple) metrics are used. Since DWQ covers a wide range of aspects, it inevitably shows some shortcomings relating to specific questions of detail, e. g. concerning sophisticated data quality metrics.

Apart from TDQM and DWQ, several minor research activities have been launched during the last two to three years, reflecting the rising awareness of the importance of DQM: In the CARAVEL project [GFSS00], a metadata-based cleansing system called AJAX has been developed, including modules for the specification, optimization, execution, and explanation of cleansing tasks. The main focus of CARAVEL lies on the elimination of duplicates (quality characteristic "absence of redundancy", cf. Fig. 1). This approach is similar to the IntelliClean project [LLL00], the Sorted-Neighborhood Method by [HS95], and related statistical methods [Jar89]. The project HiQiQ [NLF99] deals with the inclusion of data quality information into query processing in distributed databases, based on a wrapper/mediator architecture. [RCH99] describe an interactive framework for data cleansing, called Potter's Wheel A-B-C, that tightly integrates data transformation and error detection. Potter's Wheel A-B-C allows the user to gradually build transformations by adding or undoing transforms through a spreadsheet-like interface. In the background, data deficiencies are flagged as they are found. [MM00] and [DJ99] use data mining methods to detect errors automatically, comparable with the commercial tool WizRule [Wiz01].

All in all, even though there are quite a few projects dealing with data quality aspects, there is still a serious lack of formally funded methods to measure data quality. Furthermore, there is no process model supporting quality-driven data integration.

5 Conclusion

In this paper, we have proposed a quality management system for the integration of data from heterogeneous sources. Following a rigorous analogy between the integration process and a conventional manufacturing process, we were able to map the requirements of the ISO 9001:2000 standard to our model and thus implement a standard-compliant data quality management system for data warehouse systems.

During the next months, we will evaluate the DQMS by means of the epidemiological cancer registry of Lower-Saxony [HP99] which features a typical data warehouse architecture. Further future work will include the adaptation of the DQMS to the metadata standard CWM (Common Warehouse Metamodel) [OMG00].

References

[AL99] Awalt, D., Lawton, B. K.: Introduction to Data Transformation Services. In: *SQL Server Magazine*, **1** (1): 34–36, 1999.

[Ber97] Bernstein, P. A. et al.: The Microsoft Repository. In: *Proc. of the 23rd Intl. Conf. on Very Large Databases (VLDB '97), Athens, Greece*, 1997.

[BT99] Ballou, D. P., Tayi, G. K.: Enhancing Data Quality in Data Warehouse Environments. In: *Communications of the ACM*, **42** (1): 73–78, 1999.

[BW84] Basili, V. R., Weiss, D. M.: A Method for Collecting Valid Software Engineering Data. In: *IEEE Transactions on Software Engineering*, **10** (6): 728–738, 1984.

[DeM82] DeMarco, T.: *Controlling Software Projects*, Yourdon Press, New York, 1982.

[Dem86] Deming, W. E.: *Out of the crisis*, MIT, Center for Advanced Engineering Study, 1986.

[DJ99] Dasu, T., Johnson, T.: Hunting of the Snark: Finding Data Glitches with Data Mining Methods. In: *Proc. Information Quality IQ1999, MIT, Boston, MA*, 1999.

[DM92] DeLone, W. H., McLean, E. R.: Information Systems Success: The Quest for the Dependent Variable. In: *Inf. Systems Research*, **3** (1): 60–95, 1992.

[Eng99] English, L. P.: *Improving Data Warehouse and Business Information Quality*. Wiley, New York, 1999.

[GFSS00] Galhardas, H., Florescu, D., Shasha, D., Simon, E.: Declaratively Cleaning your Data using AJAX. In: *Journ. Bases de Données Avancées*, Oct. 2000.

[Hel00] Helfert, M.: Massnahmen und Konzepte zur Sicherung der Datenqualitaet. In: *Data Warehousing Strategie – Erfahrungen, Methoden, Visionen*, pp. 61–77, Springer, Berlin, 2000 (in German).

[Hin01] Hinrichs, H.: Datenqualitaetsmanagement in Data Warehouse-Umgebungen. In: *Datenbanksysteme in Buero, Technik und Wissenschaft, 9. GI-Fachtagung BTW 2001, Oldenburg*, pp. 187–206, Springer, Berlin, 2001 (in German).

[HP99] Hinrichs, H., Panienski, K.: Experiences with Knowledge-Based Data Cleansing at the Epidemiological Cancer Registry of Lower-Saxony. In: *XPS-99: Knowledge-Based Systems - Survey and Future Directions*, pp. 218–226, LNAI 1570, Springer, Berlin, 1999.

[HS95] Hernandez, M. A., Stolfo, S. J.: The Merge/Purge Problem for Large Databases. In: *Proc. of the 1995 ACM SIGMOD Conference*, 1995.

[HW00] Hinrichs, H., Wilkens, T.: Metadata-Based Data Auditing. In: *Data Mining II (Proc. of the 2nd Intl. Conf. on Data Mining, Cambridge, UK)*, pp. 141–150, WIT Press, Southampton, 2000.

[Ilo01] Ilog Inc.: *http://www.ilog.com/products/rules*, 2001.

[Inm92] Inmon, W. H.: *Building the Data Warehouse*. Wiley, New York, 1992.

[ISO00a] International Organization for Standardization: *ISO 9000:2000: Quality Management Systems – Fundamentals and Vocabulary*. Beuth, Berlin, 2000.

[ISO00b] International Organization for Standardization: *ISO 9001:2000: Quality Management Systems – Requirements*. Beuth, Berlin, 2000.

[Jar89] Jaro, M. A.: Advances in Record Linkage Methodology as Applied to Matching the 1985 Census of Tampa, Florida. In: *Journal of the American Statistical Association*, **84**: 414–420, 1989.

[JJQV98] Jarke, M., Jeusfeld, M. A., Quix, C., Vassiliadis, P.: Architecture and Quality in Data Warehouses. In: *Proc. of the 10th Intl. Conf. CAiSE*98, Pisa, Italy*, pp. 93–113, Springer, Berlin, 1998.

[Kim98] Kimball, R.: *The Data Warehouse Lifecycle Toolkit*. Wiley, N. Y., 1998.

[KSD96] Kohavi, R., Sommerfield, D., Dougherty, J.: Data Mining using MLC++ – A Machine Learning Library. In: *Tools with AI 1996*, pp. 234–245, 1996.

[LLL00] Lee, M. L., Ling, T. W., Low W. L.: IntelliClean – A Knowledge-Based Intelligent Data Cleaner. In: *Proc. of the 6th ACM SIGKDD Intl. Conf. on Knowledge Discovery and Data Mining, Boston, MA*, 2000.

[MDC01] Meta Data Coalition: *http://www.MDCinfo.com*, 2001.

[MM00] Maletic, J. I., Marcus, A.: Data Cleansing – Beyond Integrity Analysis. In: *Proc. of the Conf. on Information Quality IQ2000, MIT, Boston, MA*, pp. 200–209, 2000.

[NLF99] Naumann, F., Leser, U., Freytag, J. C.: Quality-Driven Integration of Heterogeneous Information Sources. In: *Proc. of the 1999 Intl. Conf. on Very Large Databases (VLDB '99), Edinburgh, UK*, 1999.

[OMG00] Object Management Group: *Common Warehouse Metamodel (CWM) Specification*, 2000.

[RCH99] Raman, V., Chou, A., Hellerstein, J. M.: Scalable Spreadsheets for Interactive Data Analysis. In: *Proc. of the ACM-SIGMOD Workshop on Research Issues in Data Mining and Knowledge Discovery (DMKD), Philadelphia*, 1999.

[Sha99] Shanks, G.: Semiotic Approach to Understanding Representation in Information Systems. In: *Proc. of the IS Foundations Workshop, ICS Macquarie University, Sydney*, 1999.

[She31] Shewhart, W. A.: *Economic Control of Quality of Manufactured Product*. D. Van Nostrand, New York, 1931.

[Val01] Vality Technology Inc.: *http://www.vality.com*, 2001.

[Wan98] Wang, R. Y.: A Product Perspective on Total Data Quality Management. In: *Communications of the ACM*, **41** (2): 58–65, 1998.

[Wiz01] WizSoft Inc.: *http://www.wizsoft.com*, 2001.