Institute of Information Management

University of St.Gallen

Method for Specifying Business-oriented Data Quality Metrics

Kai M. Hüner

Report No.: BE HSG / CC CDQ / 28
Chair: Prof. Dr. H. Österle
Version: 1.0
Date: March 26, 2011

University of St. Gallen for Business Administration, Economics, Law and Social Sciences (HSG)

Institute of Information Management
Müller-Friedberg-Strasse 8
CH-9000 St. Gallen, Switzerland
Tel.: +41 71 224 2420
Fax: +41 71 224 2777

Contents

Abbreviations

BE	Business Engineering
BoM	Bill of material
CC CDQ	Competence Center for Corporate Data Quality
cf.	Confer (compare)
CHF	Swiss francs
CRM	Customer Relationship Management
DAMA	Data Management Association
DQ	Data quality
DQM	Data quality management
DSS	Decision support systems
e.g.	Exempli gratia (for example)
ERP	Enterprise resource planning
EUR	Euro
GS1	Global Standards One
GTIN	Global Trade Item Number
i.e.	Id est (that is)
IEC	International Electrotechnical Commission
IEEE	Institute of Electrical and Electronics Engineers
IS	Information system
ISO	International Organization for Standardization
PLM	Product Lifecycle Management
TQdM	Total Quality data Management

Figures

Tables

Abstract

The quality of data (in particular master data) is becoming increasingly important in businesses. This is due to the increasing number of statutory and regulatory stipulations as well as the growing importance of information systems used to support decision-making and the data quality (e.g. timeliness, accuracy, completeness) these systems require. Effective data quality management calls for metrics that enable monitoring of such business-oriented data quality requirements and the identification of business-critical data defects at an early stage. The individual design of data quality metrics taking into account company-specific business processes, controlling systems and application landscapes ensures the business relevance of the metrics and enables data quality to be measured so that business-critical data defects can be identified in good time.

This paper describes a method for identifying causalities between data defects, business problems and strategic corporate objectives, and for specifying data quality metrics to monitor the identified data defects. The method follows a top-down approach and only specifies metrics for data defects which can lead to critical business problems. Validation rules are proposes as the preferred measuring technique. By implementing these rules, a measuring system (not a focus of the method) can verify data characteristics and thus identify data defects (i.e. characteristics which can lead to known business problems).

The method presented is described as the result of design-oriented research and explains various elements of the method using examples of its application in three companies. The examples illustrate the fact that it is not necessary to use all the elements of the method as the method can be configured for a specific application. To support further applications of the method, the appendix to this paper contains various examples of the requirements to be met in terms of data quality metrics, data quality dimensions, causal chains, cost types associated with defective data and validation rules which have been collected from case studies, focus group interviews and literature research.

1 Introduction

High data quality is a critical success factor for businesses when it comes to addressing increasingly tough competition in the market as well as growing regulatory requirements. The following examples illustrate the strategic importance of qualitatively good corporate data.

> *The standardization organization Global Standards One (GS1) specifies among others the Global Trade Item Identification Number (GTIN) for the unique identification of products worldwide [GS1 2010]. External service providers offer so-called data pools based on the GS1 numbering system for the multilateral synchronization of product master data [Nakatani et al. 2006].*

Retail organizations and consumer goods manufacturers benefit from the use of data pools which support the inter-organizational synchronization of master data [Schemm 2008]. However, manufacturers who make their product master data available in a pool of this kind have to mark each component of an article which is of relevance to logistics processes and ensure that they adhere to the GS1 guidelines (e.g. the extent to which the characteristics of products with the same GTIN may vary). As an example, products with the same contents (e.g. bottles of shampoo) but a weight difference which is greater than 20 percent have different GTINs. Compliance with requirements of this type is necessary to prevent problems during transportation (pallets exceeding permitted weight due to higher product weights) and at the customer's end (products not fitting the shelves due to different dimensions).

> *The provision of resource-intensive telecommunications products such as IP-TV (TV programs transmitted via the internet) calls for transparency and flexibility with regard to the utilized network infrastructure. If a subscriber places an order, a fast decision needs to be made on whether the requested product can be offered with the existing infrastructure or in what time frame the necessary infrastructure expansion can be realized. For this availability check, network infrastructure,*

> *subscriber and product data has to be correlated and analyzed in order to decide whether the line in question is capable of providing the requested product, both physically and taking into account the bandwidth sold (i.e. contractually assured).*

If the answer to an availability query is incorrect or takes too long, this can lead to customers going to another provider. Poor data quality (e.g. incomplete documentation of the network infrastructure) can therefore directly jeopardize the chances of winning new customers.

Despite the obvious strategic importance of data quality in the outlined examples, the correlation between data quality and the achievement of corporate objectives is not always easy to identify. One reason for this is the multidimensionality and context-dependence of data quality [Wang/Strong 1996]: Different data quality dimensions (DQ dimensions, e.g. timeliness, accuracy, completeness) are to be differentiated for different data object types (e.g. customer data, product data). And in order to define target levels for individual DQ dimensions, the respective usage context for the data (e.g. checking the availability of a specific bandwidth on a telecommunications line) has to be analyzed. A cost-benefit analysis is also required when looking for meaningful target levels [Lee et al. 2006]: For instance, initiatives to improve the quality of the infrastructure data described in the example should only be carried out if the costs of the initiative do not exceed the anticipated value of new customers.

The prerequisite for the cost-benefit analysis described above as well as the definition of target levels for DQ dimensions is the identification of cause-and-effect relationships (causalities) between data defects (e.g. missing infrastructure data), business problems (e.g. incorrect information on product availability) and costs (e.g. value of a new customer). Business-oriented data quality metrics (DQ metrics) can be specified on the basis of these causal chains, i.e. metrics for measuring DQ dimensions which are critical for the business success of an enterprise.

There are various examples of research work that examines data quality in general and DQ metrics in particular, and proposes concrete measuring techniques (cf. section 4). Work also exists with the requirements to be met by DQ metrics and recommendations for the design of such metrics. What is missing, however, is a structured approach to identifying the stated causalities and specifying metrics for monitoring the identified data defects. This gap gives rise to the research objective of designing a method a) for identifying causalities between data defects, business problems and strategic objectives, and b) for specifying DQ metrics for monitoring the identified data defects.

The research objective is of particular interest to enterprises with company-wide data quality management (DQM). In this case, the aim is not only to identify individual causalities such as the examples explained in the introduction, but also to apply the method repeatedly and to identify different correlated causalities. In addition to the positive effect on transport processes, an improvement in the quality of logistical product data can be expected to reduce the number of complaints and queries.

Following an explanation of fundamental concepts in section 2 and a short description of the research process in section 3, similar work is presented in section 4 and the respective results discussed with reference to the research objective of the work at hand. The method for specifying business-oriented DQ metrics is presented as the research result in section 5. Method elements are described with examples of applying the method in order to better illustrate the purpose of the individual elements. Limitations in the application of the method as well as requirements for further research are discussed in section 6.

2 Fundamental Concepts

The outlined method is the result of design-oriented research which utilizes the meta-model of the Business Engineering (BE) core as a conceptual framework. Section 2.1 therefore provides a brief overview of BE and classifies the concepts of the method according to the BE core meta-model. The goal of using DQ metrics (result of applying the method) is to evaluate data quality objectively and to enable comparable data quality measurements (DQ measurements) as a preventive DQM initiative. For this reason, the understanding of data quality and DQM on which the work at hand is based is discussed briefly in section 2.2.

2.1 Business Engineering as methodological framework

BE is a systematic approach which supports enterprises in the implementation of transformation projects (e.g. strategic realignment, design of strategy-compliant business processes). Rather than focusing on isolated aspects of corporate transformation (e.g. strategy development, organizational design), BE structures a complex design task on various modeling layers by means of manageable activities with defined results and relationships [Österle 1995, Scheer 1998].

The St. Gallen approach to BE [Österle 1995, Österle/Blessing 2003, Österle/Blessing 2005] is the methodological framework for this paper. The approach structures transformation projects on the modeling layers *strategy*, *process* and *system* by means of methods (cf. section 3.2), and specifies design objects of the result documents (results of applying a method) in a meta-model. HÖNING [2009] identifies the core (i.e. similar elements) of various BE approaches, structures it using method engineering concepts (e.g. activities, roles, techniques) and specifies the design objects of the BE core in a meta-model [Österle et al. 2007, Weber 2009].

Figure 2-1 classifies the concepts of the presented method according to the BE core meta-model, taking into account the stipulated extension rules [Höning 2009]. The concepts marked with the prefix *BE::* are meta-entity types and relations of the BE core; extended meta-entity types are shaded gray. The modeling of the extended concepts is oriented toward the meta-model of the master data integration method used by SCHMIDT [2010].

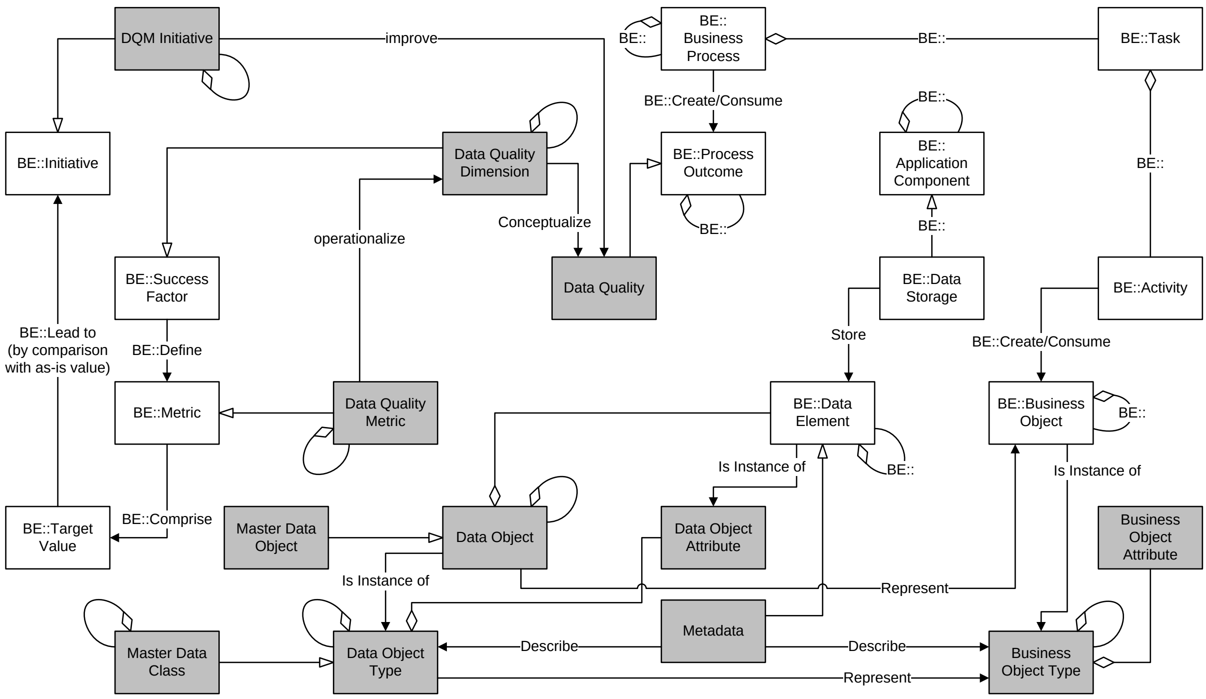


Figure 2-1: Classification of the Concepts of the Method According to BE Core

Application of the method is not restricted to master data, but is fundamentally possible for all data elements, data object attributes and metadata whose defects cause business problems. If no differentiation is required (e.g. for specific DQ dimensions in the appendix or concrete examples in section 5.2), the general identifier *data* is used in the following. However, the majority of applications of the method implemented to date have identified data defects of data elements which the companies refer to as master data. The differentiation of different data types such as transaction data, metadata and master data has been dealt with in other works [Burnett et al. 1999, Wedekind 2001, Mertens 2007, Loshin 2008, Smith/McKeen 2008].

Section 5.1 describes the concrete meta-model of the presented method as a specification of the design objects of the documentation model and structures the concepts shown in Figure 2-1 using result documents. Figure 2-1 merely shows the relationships of central concepts of the presented method to the design objects of the BE core. Appendix A describes the extended meta-entity types; the design object types of the BE core are specified by HÖNING [2009].

2.2 Data quality and data quality management

Data quality is the object of various research works with different research approaches. Some of the works propose lists and categories with DQ dimensions as the result of empirical research [Wang/Strong 1996], while others describe experience from professional and consultancy practice [Redman 1996, English 1999, Loshin 2001], or present DQ dimensions as the result of theoretical conclusions [Wand/Wang 1996, Price/Shanks 2005].

DQM encompasses initiatives aimed at improving the quality of data [Batini/Scannapieco 2006], but goes beyond a purely reactive improvement of data quality (i.e. identification and elimination of data defects) [Shankaranarayanan/Cai 2006]. In fact, DQM consists of the proactive and preventive improvement of data quality through a continuous cycle of definition, measurement, analysis and improvement of data quality as well as the design of the required framework [Wang et al. 1998, English 1999, Eppler/Helfert 2004]. An overview of the various approaches for DQM is provided by BATINI ET AL. [2009].

3 Research Methodology

The description of the process by which the presented method is created as an artifact [Simon 1996] enables the method to be reproduced and evaluated as a result of scientific research. It should be borne in mind that this paper focuses on the design goal of information systems research, rather than on the epistemological goal [Frank 1997, Becker et al. 2003]: The research objective is to design a method for the specification of business-oriented DQ metrics. A gain in scientific knowledge will of course be necessary in order to achieve this objective, e.g. identification of the correlations between data defects, business problems and DQ metrics described in the meta-model of the method (cf. section 5.1) [vom Brocke 2003]. However, the main emphasis is on the method as the research result and on the business benefits derived from its application. Section 5.2 describes the results of applying the method in three participative case studies [Baskerville 1997] which show that application of the method can produce useful results with acceptable effort.

3.1 Research process

The presented method is the result of design-oriented research which follows Design Science Research as the research methodology paradigm. The aim of Design Science Research is to design artifacts as a solution to practical problems [March/Smith 1995, Hevner et al. 2004]. Proposed artifact types are constructs, models, methods (type of artifact presented, cf. section 3.2) and instantiations.

The Competence Center for Corporate Data Quality[1] (CC CDQ) provides the research context for the method design process. The CC CDQ is part of the BE research program[2] at the University of St. Gallen. Since 2006, it has been developing solutions to support quality management for data used on an enterprise-wide basis in collaboration with various organiza-

[1] http://cdq.iwi.unisg.ch
[2] http://www.iwi.unisg/behsg/

tions[3] under the auspices of the Institute for Information Management[4] [Österle/Otto 2010, Otto/Österle 2010].

In accordance with the various requirements to be met by the design of artifacts [Hevner et al. 2004], the process for designing an artifact consists of the phases problem identification, definition of design goals, designing artifacts to resolve a problem and evaluation of the artifacts [Gregor 2006, Peffers et al. 2008]. Focus group interviews [Morgan/Krueger 1993, Rosemann/Vessey 2008] with business experts from different organizations in the CC CDQ and participative case studies [Baskerville 1997] have supported the phases problem identification, goal definition and evaluation of the design process of the presented method. Examples of applying the method in participative case studies explain individual elements of the method in section 5.2. Evaluation of the method in a focus group interview is explained in section 6.

3.2 Method engineering

In method engineering, a method denotes a systematic approach to information systems (IS) development [Heym/Österle 1993]. Method engineering is the disciplined process of building, improving or modifying a method by means of specifying the method's components and their relationships in a methodological (i.e. engineered) way [Heym/Österle 1993, Gutzwiller 1994]. A method contains a product description and a process description [Olle et al. 1988, Heym/Österle 1993, Brinkkemper 1996]. The process elements guide the design of the method results: They define which elements of the meta-model are to be instantiated in what order and who delivers the information required for the instantiation or performs the respective activities.

[3] http://cdq.iwi.unisg.ch/index.php?id=144
[4] http://www.iwi.unisg

The presented method has been designed as the instantiation of a meta-model (see Figure 3-1) proposed by GUTZWILLER [1994] that defines the following components of a method. A comparable structure is described by NUSEIBEH ET AL. [1996].

- *Procedure model*. An activity is a task aimed at the design of one or more defined results. Activities can be hierarchically structured and can belong to a sequence. All activities of a method form the procedure model of the method.
- *Roles*. People or boards who exercise a specific role and perform activities. From the viewpoint of a sponsor, a role groups together various activities. Roles are always involved in activities in a specific form (e.g. *responsible* or *consulting*).
- *Documentation model*. Activities utilize results and create or modify results. Like activities, results can be hierarchically structured. All results together form the documentation model of the method.
- *Meta-model*. The meta-model specifies the design objects of the results. It defines the syntax and semantics of the results in the sense of a formal language.
- *Techniques*. Techniques describe how results are to be produced. Unlike the procedure model, which defines *when* and in *what order* results are to be created, techniques describe *how* results are to be obtained.

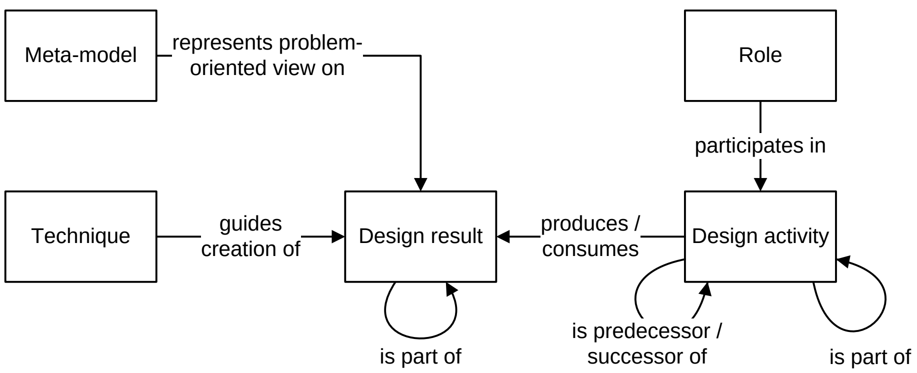


Figure 3-1: Meta-Model of Method Engineering [Gutzwiller 1994]

4 State of the Art

In addition to the differentiation of various DQ dimensions (cf. section 2.2), DQ measurements, in other words the evaluation of individual dimensions by means of objectively determined metrics, have also been the object of various research works. The following sections present approaches which describe concrete DQ metrics, measuring techniques, procedures for identifying data defects or calculation formulas for DQ metrics.

Some of the articles discussed distinguish between the concepts of *data* and *information,* and hence *data* and *information quality*. The presented method considers *data* as the representation of business objects stored in databases and utilized by data users (cf. section 2.1). As in other works with a similar focus, no distinction is made here between *data* and *information* [Wang et al. 1998, Bourdreau/Couillard 1999, Pipino et al. 2002, McKinney Jr./Yoos 2010]. Other authors provide a nuanced discussion of the distinction between these concepts [Bourdreau/Couillard 1999, Spiegler 2000, Boisot/Canals 2004, McKinney Jr./Yoos 2010].

4.1 Batini et al.

BATINI/SCANNAPIECO [2006] describe various DQ dimensions and explain these using calculation formulas. Furthermore, they outline various cost-benefit classifications and explain the necessity of identifying business-critical data defects and monitoring them with DQ metrics . A comparative analysis of various approaches to measuring and improving data quality is explained by BATINI ET AL. [2009]. BATINI ET AL. [2007] describe a methodological procedure with the following four phases to identify and monitor data defects (*data quality risk*) in the data objects of an enterprise:

- *Data quality risk prioritization*. Modeling of organization units, processes and utilized data as well as the relationships for which data defects are known or assumed.

- *Data quality risk identification*. Identification of concrete business problems where data defects are the cause.
- *Data quality risk measurement*. Qualitative and quantitative measurement of data quality (not further differentiated) and evaluation of the impact of data defects.
- *Data quality risk monitoring*. Periodical DQ measurement, i.e. monitoring and evaluation of identified data defects.

4.2 Caballero et al.

CABALLERO ET AL. [2007] define a meta-model for DQ measurements and metrics that is based on ISO/IEC Standard 15939 (cf. section 4.7). In another work, CABALLERO ET AL. [2008] use the meta-model to describe a methodology for DQ measurements that outlines a procedure for establishing DQ measurements, planning a DQ measuring process, performing DQ measurements and evaluating existing measuring processes in four phases with 16 activities. The four phases consist of the following:

- *Establish and sustain the data quality measurement commitment*. Setting up a group of employees responsible for the specification, use and maintenance of DQ metrics.
- *Plan the data quality measurement process*. Identifying business problems, affected process activities, data and roles, identifying and prioritizing requirements to be met by utilized data, identifying measurable data characteristics, specifying DQ metrics for monitoring characteristics and selecting the data objects to be measured.
- *Perform the data quality measurement process*. Integrating processes for DQ measurements into existing DQM processes, specifying data quality reports for communicating measurement results, performing DQ measurements and communicating measurement results.
- *Evaluate data quality measurement process*. Continuously improving DQ metrics and measurements.

4.3 DAMA

The Data Management Association (DAMA) defines nine functions for data management [DAMA 2009] and in the DQM function describes the definitions of DQ metrics and the definition of validation rules (*data quality business rules*) as two of twelve activities. The definition of DQ metrics consists of the following six steps:

- Select one of the previously identified business problems.
- Identify associated data object types, data utilization processes and data maintenance processes.
- List the requirements to be met by the identified data [DAMA 2009].
- Identify impaired DQ dimensions for each requirement and specify validation rules in order verify if data objects meet the requirements.
- Describe the measuring process for each validation rule, i.e. how the characteristics of a data object defined by the rules are to be validated.
- Specify a threshold value for each validation rule.

4.4 Larry P. English

ENGLISH [1999] defines data quality (*information quality*) as "consistently meeting knowledge worker and end-customer expectations" and, in so doing, extends the context-dependence of data quality outlined by WANG/STRONG [Wang/Strong 1996] to include the element of consistent data provision. Here, consistency encompasses a) fulfilling the requirements of all (not just specific) data users, and b) providing the same data (also from different databases) for the same business objects.

Furthermore, ENGLISH [1999] defines information as the aggregate of data, the clear meaning of data (definition) and comprehensible presentation, and states different DQ dimensions for the three components [English 1999]. This differentiation is also to be found in the proposed methodology, *Total Quality data Management* (TQdM), for sustainable improvement of data quality [English 1999], which distinguishes between processes for measuring data definition quality (cf. Figure 4-1, P1) and data quality (P2). To measure the impact of poor data quality (*nonquality information costs*, P3), ENGLISH [1999] defines various cost types for defective data. Figure 4-2 shows the individual activities of the processes for DQ measurements.

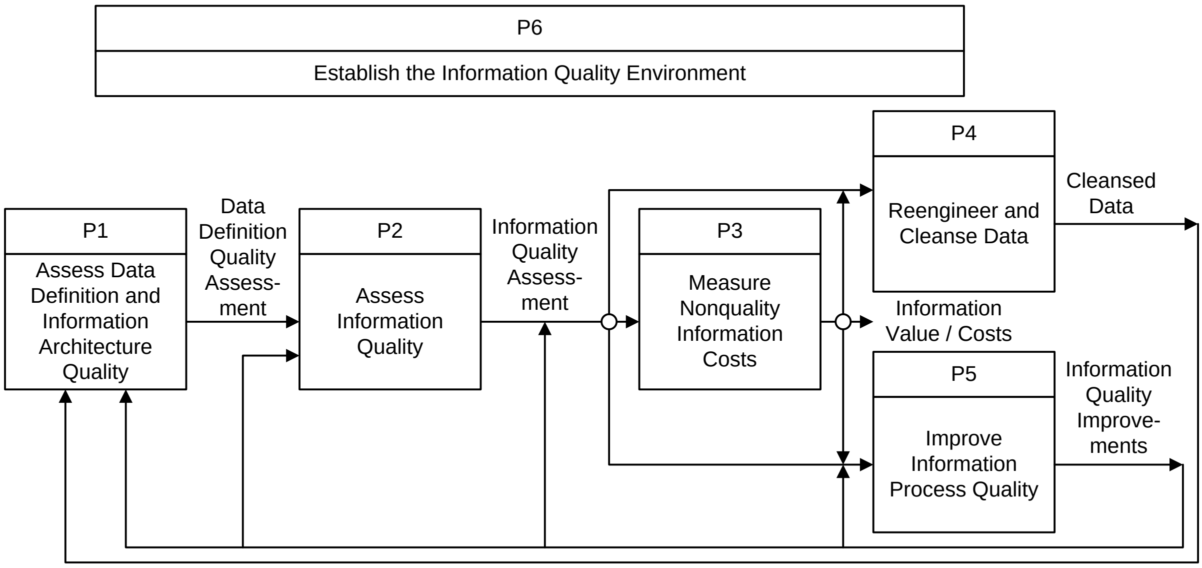


Figure 4-1: Processes of the TQdM Methodology [English 1999]

As data definitions and information architecture (for the sake of simplicity, *metadata* in the following) form the basis for specific DQ measurements in P2, individual activities of P1 are to be performed before the start of P2 (cf. Figure 4-2, S1.1, S1.2 and S2.1). Alongside the activities for the actual DQ measurements (S1.4, S1.5, S1.6 and S2.7), the measuring processes include activities for limiting a scope (S1.2, S2.1), identifying stakeholders (S1.3), specifying target levels and metrics (S2.2) as well as comparing costs and benefits of data storage and use [English 1999].

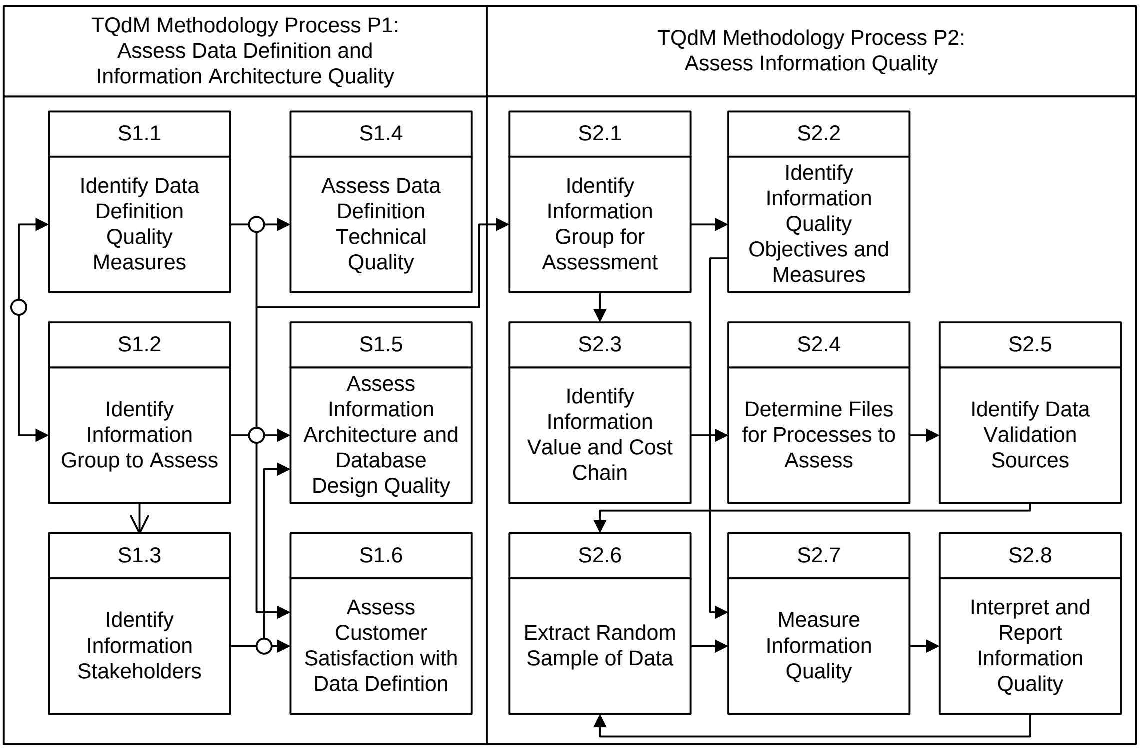


Figure 4-2: Process for Measuring Data Quality [English 1999]

4.5 Heinrich et al.

KAISER ET AL. [2007] define six requirements to be met by DQ metrics (also described by HEINRICH ET AL. [2007, 2008]) as well as metrics for the dimensions *timeliness* (or *currency*) [Heinrich et al. 2007, Heinrich et al. 2009], *correctness* and *completeness* [Heinrich et al. 2008].

4.6 IBM

ALUR ET AL. [2007] describe a methodological procedure for identifying data defects. In the first phase (*data profiling*), frequency distributions for the same data elements are calculated, primary and foreign key integrity is analyzed, and the consistency of data types and formats is validated. The results of this phase (e.g. a list of unexpected states of data elements, a list of necessary data type corrections, primary and secondary key lists) serve as input for the second phase (*rule validation*) in which validation rules are defined and verified.

4.7 ISO / IEEE

Standard 15939 of the International Organization for Standardization and the International Electrotechnical Commission (ISO/IEC) [DIN/EN/ISO 2000] defines a measuring process for system and software development (cf. section 4.2 for adaptation to DQ measurements). The process model encompasses activities for specifying the required information for measurements, for performing measurements and for interpreting the measuring results. In addition to the process model, the standard contains a glossary of employed terms (cf. section 4.2, adaptation of terminology for DQ measurements).

Standard 1061 of the Institute of Electrical and Electronics Engineers (IEEE) [1998] defines a methodological procedure for identifying, implementing, analyzing and validating quality metrics for software and software development processes. In addition to the methodology itself, the standard defines numerous terms required to describe the methodology and based on the IEEE glossary for software engineering [IEEE 1990]. These two IEEE standards provide the conceptual basis for the above mentioned ISO/IEC standard.

4.8 David Loshin

LOSHIN [2001] defines data quality in accordance with the definition of WANG/STRONG [1996] as “fitness for use” and refines this concept with the aspects *free of defects* and *possesses desirable characteristics* as well as various DQ dimensions which are structured according to five categories [Loshin 2001]. For DQ measurements, LOSHIN [2001] describes nine categories with validation rules (*data quality rule*) as well as various examples for measuring the individual DQ dimensions [Loshin 2001].

Furthermore, LOSHIN [2001] describes a methodological procedure for identifying the costs of business problems attributable to data defects. The procedure is oriented toward data utilization processes (*information chain*) and includes the following activities:

- *Model data utilization processes*. Modeling process activities and data flows (e.g. the business process of *invoice dispatch* with data maintenance performed by customer service, updating of the invoice items by accounting management and the sending of invoices by the postal service provider).
- *Interview employees*. Identifying business problems in interviews with employees for each activity in the data utilization process.
- *Interview customers*. Identifying data defects in interviews with customers for each activity where customers utilize data (e.g. invoice, marketing activity).
- *Isolate data defects*. Annotating the data utilization process with identified data defects.
- *Identify impacts*. Assigning business problems to the individual data defect/activity pairs.
- *Evaluate business problems*. Evaluating the individual activity/business problem pairs by means of incurred costs or company-specific metrics for business problems.

To support the methodology, LOSHIN [2001] outlines various business problems caused by data defects.

4.9 Jack E. Olson

OLSON [2003] defines data quality as the fulfillment of task-specific requirements ("data has quality if it satisfies the requirements of its intended use") and refines the DQ dimension *accuracy* ("whether the data values stored for an object are the correct values") by means of six further dimensions [Olson 2003] that make a fundamental distinction between *wrong data element* and *wrong data format*. Various aspects are taken into consideration here, such as missing, inconsistent or inaccurate data elements and wrong data formats, which other authors describe by dimensions such as *consistency*, *accuracy* or *completeness*. However, the aim of this differentiation is not to provide a new definition of data quality, but to distinguish between measurable and non-measurable data defects (cf. Figure 4-3).

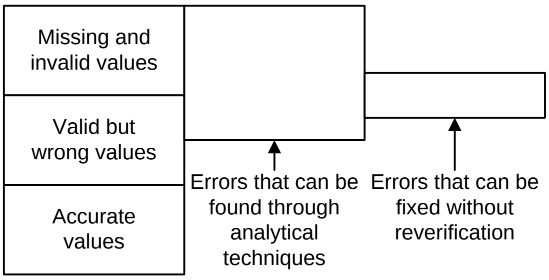


Figure 4-3: Measurable Data Defects [Olson 2003]

For measuring data quality (*data profiling*), OLSON [2003] describes five different techniques, each of which addresses different data defect types. All techniques formulate possible data defects as rules and identify by means of measurement the data objects that do not satisfy specific rules (cf. Figure 4-4).

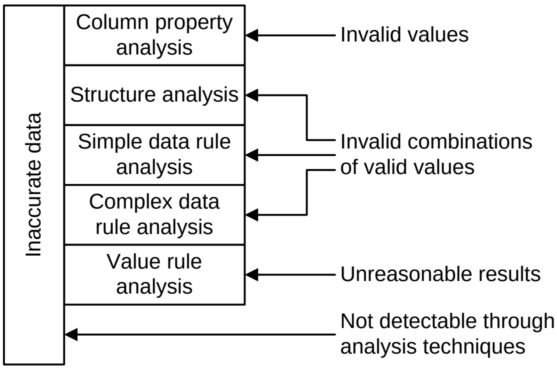


Figure 4-4: Rule-Based Technique for DQ Measurements [Olson 2003]

In addition to the detailed description of rule-based measuring techniques, OLSON [2003] describes different cost types and benefit potentials of improved data quality. He emphasizes the necessity of preparatory projects for identifying and evaluating critical data defects (*assessment projects*). While these do not have any immediate benefit, they do, however, increase the benefit of subsequent projects with concrete improvement initiatives (*implementation projects*) through the selection of effective initiatives and the prioritization of defects [Olson 2003].

4.10 Price et al.

PRICE/SHANKS [2005] conceptualize data quality from a semiotic viewpoint and define the following three categories for DQ dimensions:

- *Syntactic data quality*. These dimensions evaluate the conformity of data elements and their specification.
- *Semantic data quality*. These dimensions evaluate the representation of the real world (e.g. business objects) by data objects.
- *Pragmatic data quality*. These dimensions evaluate the fulfillment of expectations that a specific data user associates with data when processing a specific task (subjective data quality).

For measuring subjective data quality, PRICE ET AL. [2008] define 34 statements and describe their use as measuring techniques.

4.11 Thomas C. Redman

REDMAN [1996] defines data quality by means of DQ dimensions divided into three categories for measuring the quality of metadata, data elements and data presentation. He focuses on the DQ dimension *accuracy* [Redman 1996, 2005], which he defines as the ratio of the number of correct data elements to the total number of all data elements [Redman 1996].

4.12 Wang et al.

WAND/WANG [1996] define four data defect types (*data deficiencies*) and the four intrinsic DQ dimensions *complete*, *unambiguous*, *meaningful* and *correct*. Here, a data defect is a variance between the user's immediate perception of reality ("user's direct view of a real-world system") and his/her perception of an IS (e.g. data) as a representation of reality ("user's view of a real-world system as inferred from an IS representing the real-world system") [Wand/Wang 1996].

- *Incomplete representation*. The IS is not capable of representing a specific status of the real-world system.
- *Ambiguous representation*. One status of the IS represents several statuses of the real-world system.
- *Meaningless states*. One status of the IS represents an impossible status of the real-world system.
- *Garbling*. The status of an IS generated as the representation as a status of the real-world system is incorrect, i.e. the status of the real-world system is represented by a status of the IS which in the event of correct representation would represent a different status of the real-world system.

WANG/STRONG [1996] identified 15 DQ dimensions and a structure comprising four categories in a two-stage survey (identification of possible dimensions, evaluation of the dimensions collected) of data users. By considering data as a product and defining data quality as "the data's fitness for use" (cf. section 2.2), the authors shape the understanding of data quality on which the paper is based. This perspective is concretized by further articles [Strong et al. 1997, Wang 1998, Wang et al. 1998].

BALLOU ET AL. [1998] pick up on the consideration of data as a product and define a meta-model for data production systems (*information manufacturing systems*), i.e. a process model for describing data lifecycles (specification, use and maintenance). Taking the example of a DQ metric for the DQ dimension *timeliness*, BALLOU ET AL. [1998] explain how the DQ metrics values for the various meta-model elements are to be calculated. SHANKARANARAYANAN ET AL. [2000, 2003] develop this approach further.

HUANG ET AL. [1999] describe the necessity for developing company-specific DQ metrics and define the following three categories of metrics (no new metrics in comparison with WANG/STRONG [1996]).

- *Subjective DQ metrics*. Metrics for measuring data quality as perceived by data users.
- *Objective application-independent DQ metrics*. Metrics for measuring e.g. integrity or completeness.
- *Application-specific DQ metrics*. Metrics for verifying e.g. whether the requirements to be met by a data-using application are fulfilled.

To support the measurements, HUANG ET AL. [1999] describe two applications for measuring objective and application-specific metrics and a list of questions for measuring subjective DQ metrics.

LEE ET AL. [2002] define 65 statements for DQ measurements in organizations along the DQ dimensions of WANG/STRONG [1996]. The questionnaire has been successfully used in five organizations among a total of 261 people (four-page printout, questions not organized per dimension but mixed, 11-point Likert scale per question with level of agreement ranging from “not at all” to “completely”).

PIPINO ET AL. [2002] define the following three categories for calculating the values of DQ metrics and assign the 15 DQ dimensions of WANG/STRONG [1996] by way of example.

- *Simple ratio*. Ratio of the number of required results to the total number of results. Assigned dimensions: *Free of error*, *completeness*, *consistency*, *concise representation*, *relevancy* and *ease of manipulation*.
- *Min or max operation*. Minimum or maximum value of metrics values for different attributes or data objects. Assigned dimensions: *Believability* (min), *appropriate amount of data* (min), *timeliness* (max) and *accessibility* (max).

- *Weighted average*. Weighted average of metrics values for different attributes or data objects. Applicable to the dimensions stated for min/max operation.

PIPINO ET AL. [2005] outline calculation formulas (*measurement scales*) for measuring the DQ dimensions *completeness*, *correctness* and *timeliness*/*currency*. LEE ET AL. [2006] also explain various calculation formulas for measuring DQ dimensions and describe how to use a list of questions [Lee et al. 2002] for evaluating data quality.

4.13 Comparative evaluation of the presented approaches

The presented method should support a) the identification of causalities between data defects, business problems and strategic corporate objectives, and b) the specification of DQ metrics for monitoring the identified data defects (cf. section 1). The following criteria refine this research objective and enable the presented approaches to be assessed in comparison with one another as well as in respect of the research objective.

- *Method structure (K1 – K5)*. The aim of the method structure explained in section 3.2 is to reduce the complexity of a design task (here: specifying business-oriented DQ metrics) by structuring that task. Even if none of the presented approaches describes a method that is complete in terms of the outlined structure, individual activities (K1), results (K2), meta-model elements (K3), roles (K4) and techniques (K5) are comparable and have supported the method design.
- *Pragmatic definition and presentation (K6)*. The aim of applying the method is to specify metrics for monitoring business-critical data defects. This assumes the identification of business-critical data defects for which business data users are important discussion partners. In order to be able to explain data defects and possible impacts in interviews or workshops, a comprehensible and problem-oriented description of the phenomenon *data quality* is necessary.

- *Metrics with business relevance (K7).* The aim is to design business-oriented DQ metrics, i.e. metrics for monitoring data defects that can cause critical business problems. The design process should therefore include activities that identify and evaluate causalities between data defects and business problems.
- *Implementable metrics (K8).* The aim is a detailed documentation of the DQ metrics so that a prototype implementation of the measuring techniques and interpretation of the metrics values on a measurement scale are possible. In particular, this calls for the concrete description of a measuring technique (e.g. questionnaire, validation rules).

Article	(K1) Procedure	(K2) Documentation Model	(K3) Meta-Model	(K4) Roles	(K5) Techniques	(K6) Definition and Presentation	(K7) Business Relevance	(K8) Implement-ability
Alur et al. [2007]	◔	○	○	○	●	○	○	○
Ballou et al. [1998]	○	○	○	○	◕	○	○	◑
Batini et al. [2007]	◑	○	○	○	○	○	◑	○
Batini et al. [2009]	◔	○	○	○	○	○	◔	○
Batini/Scannapieco [2006]	○	○	○	○	◔	◔	◔	◔
Caballero et al. [2007]	○	○	◕	○	○	○	○	○
Caballero et al. [2008]	◕	◔	○	○	◔	○	○	○
DAMA [2009]	◑	○	○	○	○	◔	○	○
English [1999]	◕	○	○	◑	◔	◕	◑	○
Heinrich et al. [2009]	◑	○	○	○	○	○	○	◑
Huang et al. [1999]	○	○	○	○	◕	◑	◑	◕
IEEE [1990]	○	○	◔	○	○	○	○	○
IEEE [1998]	○	○	◑	○	○	○	○	○
ISO/IEC [2007]	○	○	◕	○	○	○	○	○
Kaiser et al. [2007]	○	○	○	○	◑	○	○	◔
Lee et al. [2002]	◔	○	○	○	◕	○	○	◕
Lee et al. [2006]	○	○	○	○	○	◔	○	◑
Loshin [2001]	◑	○	○	○	○	◕	◕	◑
Olson [2003]	◔	○	○	○	◕	◕	◑	◑

Article	(K1) Procedure	(K2) Documentation Model	(K3) Meta-Model	(K4) Roles	(K5) Techniques	(K6) Definition and Presentation	(K7) Business Relevance	(K8) Implementability
Pipino et al. [2002]	○	○	○	○	◑	◑	○	◔
Pipino et al. [2005]	○	○	○	○	◑	◔	○	◕
Price et al. [2008]	○	○	○	○	◕	○	○	◕
Price/Shanks [2005]	○	○	○	○	○	◕	○	○
Redman [1996]	○	○	○	○	○	◑	○	◔
Wand/Wang [1996]	○	○	○	○	○	◕	○	○
Wang/Strong [1996]	○	○	○	○	○	●	○	○

○: Approach/result does not consider respective aspect, ◔: Approach/result considers respective aspect, ◑: Approach/result is suitable as template for designing the method, ◕: Approach/result is suitable for use in the method with modifications, ●: Approach/result is used in the method.

Table 4-1: Evaluation of Presented Approaches to Data Quality Measurement

Table 4-1 shows the evaluation of the presented approaches and results along eight criteria on a five-point ordinal scale. The maximum evaluation means that the respective approach and/or the respective result have been reused when designing the method without any major adjustments.

5 Method Design and Application of the Method

This section explains the individual elements of the method for specifying business-oriented DQ metrics. Since procedure model, documentation model, roles and techniques are closely linked with one another, these partial models are not described individually but in the form of method fragments (in each case, design result, activity producing the result, roles involved and techniques used) in sections 5.2.1 to 5.2.7 [Bucher 2009]. Techniques and roles are described when the respective role or technique is involved in the execution of an activity for the first time or can be used for this purpose.

5.1 Meta-model

The aim of applying the presented method is to

- identify causal chains (i.e. of associated cause-and-effect relationships) between data defects, critical business problems and strategic corporate objectives, and to
- specify DQ metrics for monitoring the identified data defects.

Figure 5-1 illustrates the concept of causal chains by classifying the concepts according to the various layers of an enterprise architecture. This simplified presentation shows the relationship between activities of a company's business processes and the achievement of strategic business objectives as well as the utilization of data when creating products and services (cf. Figure 5-1, left). For the identification of concrete causal chains (cf. Figure 5-1, right), all relationships between strategic objectives, process activities and utilized data are to be considered and the chains are to be selected

- which contain a business problem that jeopardizes the achievement of a strategic corporate objective and
- whose business problems are attributable to clearly identifiable data defects.

A metric which monitors a data defect identified by this means, has a clearly defined relationship to the company's value creation via the causal chain and is therefore a business-oriented DQ metric.

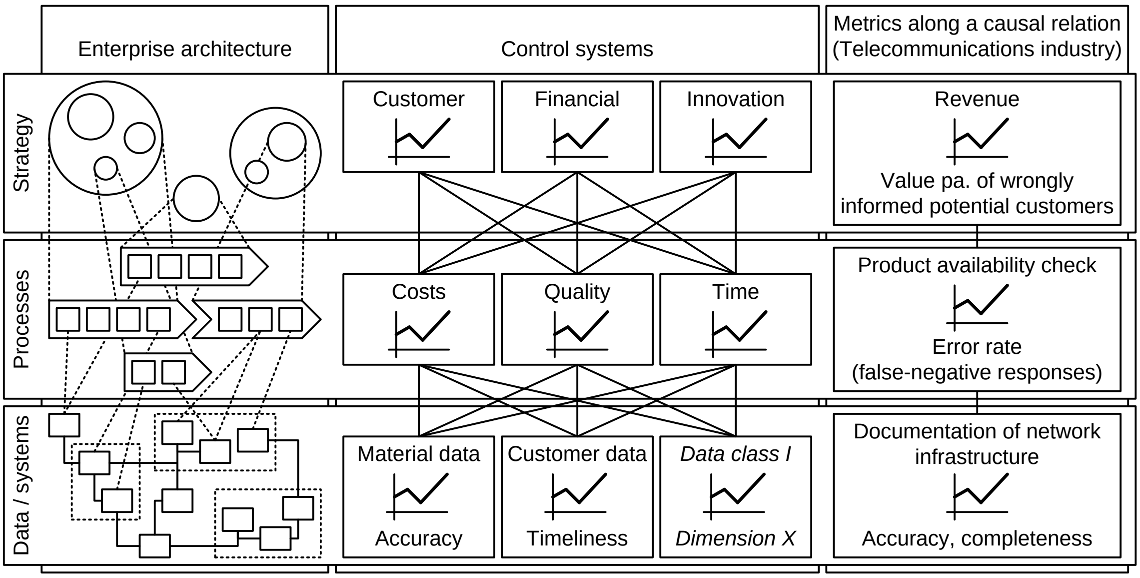


Figure 5-1: Causal Chains in the Context of an Enterprise Architecture

As strategic success factors are usually monitored by process metrics or strategic controlling systems (e.g. Balanced Scorecard [Kaplan/Norton 1992]) (cf. Figure 5-1, center), these monitoring structures are to be considered both when identifying the causal chains and when specifying DQ metrics (cf. section 5.2.5).

Figure 5-2 shows the meta-model of the method with attributes and relationships of the design objects of the documentation model. Key design objects are data defects and DQ metrics. The remaining design objects are for documenting the identification process and the business relevance of the identified data defects. The significance of these business objects is explained in the context of the individual method fragments (cf. sections 5.2.1 – 5.2.7).

The definition of the concept *data defect* is based on the definition of data quality explained in section 2.2, according to which the quality of data is considered to be good if it satisfies the expectations of a data user when executing a task. In line with this concept, a data defect is a characteristic of the data object, the data object type or the presentation of data elements which does not satisfy the respective requirements. If, for example, the process activity *invoice dispatch* expects complete address data for posting invoices, a missing street name is a data defect which reduces process performance (e.g. measured by the proportion of undeliverable invoices) and leads to a variance in the measured values of a process metric from its target level.

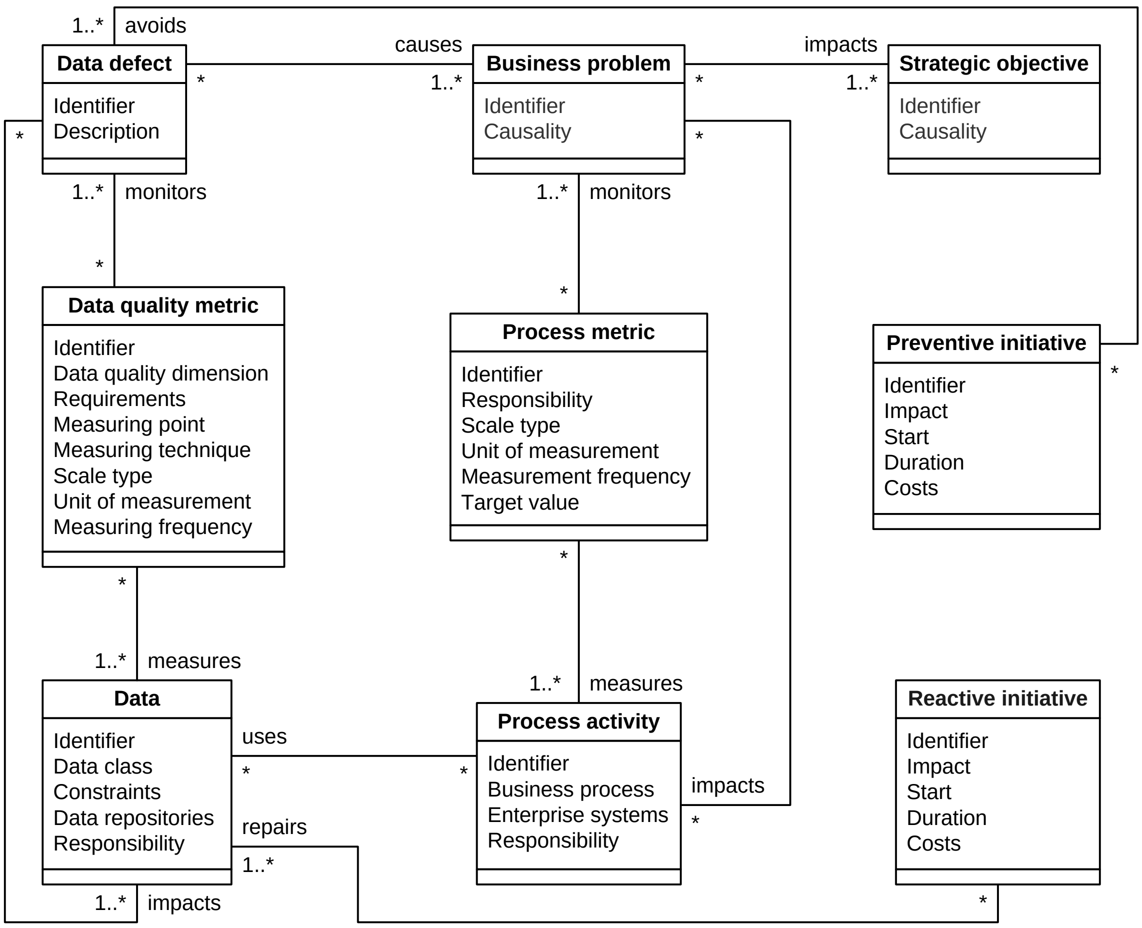


Figure 5-2: Meta-Model of the Method

The task of a DQ metric is to monitor data defects, i.e. to identify the characteristics of data objects, data object types or the presentation of data elements which do not meet the requirements at an early stage (i.e. before they cause a business problem). For the specification of DQ metrics, therefore, the requirements to be met by data are to be identified and translated into (as far as possible machine-executable) measuring techniques. For this purpose, the presented method describes a technique for formalizing requirements by means of validation rules (cf. section 5.2.5.1).

For the implementation of DQ metrics, the measuring technique and the measuring scale are particularly important. The measuring technique measures a metric value by mapping the result of a measurement performed at a measuring point onto a scale. In order to be able to compare the measured values of a data quality metric with those of a process metric (e.g. to show any negative influence of poor data quality on process performance), the level of measurement of the process metric must be taken into account when specifying the data quality metric: Comparative statements (e.g. correlations) are only permissible for the same levels of measurement.

5.2 Method fragments

The description of the presented method by means of method fragments is structured according to seven activities in a procedure model. Since procedure model, documentation model, roles and techniques are closely linked with one another, these partial models are not described individually but in the form of method fragments (in each case, design result, activity producing the result, roles involved and techniques used) in sections 5.2.1 to 5.2.7 [Bucher 2009]. Figure 5-3 provides an overview of the method with the activities (left-hand side) and the respective roles involved (right-hand side). The middle column of Figure 5-3 shows the attributes of the meta-model which instantiate the stated roles in the respective activity and thus specify the result of the individual activities. The “*” symbol means that all the attributes of the stated meta-model entity type are instantiated. The fundamental structure of the method consists of the following three phases:

- *Collect information*. The knowledge required to identify causal chains is distributed across a large number of roles in the enterprise. This phase therefore encompasses activities with a high degree of communication. Information is collected from different roles and consolidated into a scope for application of the method and into causal chains.

Activities	Results	Roles: Corporate Data Steward	Sponsor	Process Owner	Data User	Business Data Steward	Technical Data Steward
Phase I: Collect information							
I.1: Define the scope	ProcessActivity.Identifier	x	x	x			
	ProcessActivity.BusinessProcess	x	x	x			
	ProcessActivity.Responsibility	x		x			
	ProcessMetric.*	x		x			
	BusinessProblem.*	x		x			
	StrategicObjective.*	x		x			
I.2: Identify data and IT systems	ProcessActivity.EnterpriseSystem	x		x			
	Data.*	x		x			
I.3: Identify causal chains	StrategicObjective.Causality	x		x			
	BusinessProblem.Causality	x			x	x	x
	DataDefect.*	x			x	x	x
	PreventiveInitiative.*	x			x	x	x
	ReactiveInitiative.*	x			x	x	x
Phase II: Analyze and specify							
II.1: Define and prioritize requirements	DataQualityMetric.Requirements	x	x			x	x
II.2: Specify data quality metrics	Data.Constraints	x				x	x
	DataQualityMetric.Identifier	x				x	x
	DataQualityMetric.DataQualityDimension	x				x	x
	DataQualityMetric.MeasuringPoint	x					x
	DataQualityMetric.MeasuringProcedure	x					x
	DataQualityMetric.ScaleType	x				x	x
	DataQualityMetric.UnitOfMeasurement	x				x	x
	DataQualityMetric.MeasuringFrequency	x				x	x
Phase III: Approve and document							
III.1: Check the fulfillment of requirements		x				x	x
III.2: Document metrics specifications		x	x				

Figure 5-3: Procedure Model with Relationships to Meta-Model and Roles

- *Analyze and specify*. The aim is to specify all the information required to implement at least one DQ metric. The prime task is to formalize identified data defects by means of validation rules (or another measuring technique) and link the rules (e.g. through calculation formulas) to form metrics.

- *Approve and document*. The aim is to document the specifications of the DQ metrics and the connection to business (i.e. the causal chains) so that the metrics can be implemented.

As well as the description of the method, sections 5.2.1 to 5.2.7 contain examples of the instantiation and/or use of individual method elements in three different companies. Table 5-1 outlines the anonymized companies and the respective project context.

	Consumer Goods Inc.	**Power Inc.**	**Telephone Inc.**
Sector	Consumer goods production	Power supply	Telecommunications
Revenues (2009)	EUR 6 billion	CHF 750 million	EUR 60 billion
Headcount (2009)	20,000	1,400	260,000
Project Context			
Time Frame	01/2010 – 06/2010	11/2009 – 06/2010	06/2009 – 02/2010
Project Team	5 team members 17 interviewees	4 team members 9 workshop participants	3 team members 2 interviewees
Business Processes	Distribution, marketing, quality management, production	Invoicing, reminders, invoice dispatch, marketing	Availability check for broadband internet
Data Object Type	Product data	Business partner, contract, contract account data	Network infrastructure, customer data

Table 5-1: Application Companies (anonymized) and Project Context

5.2.1 Define scope

The goal of Activity I.1 is to define possible scopes (cf. Table 5-2).

<table>
<tr><td colspan="3">Activity I.1: Define Scope
The goal is to identify process activities or business transactions where business problems are known or presumed, the cause of which is presumed to be data defects. For each process activity, at least one process metric should be identified for which measured values are available. In addition, the (monetary) impact of the business problems is to be documented.</td></tr>
<tr><td>Result Document
List of scopes</td><td>Roles
• Sponsor
• Corporate data steward
• Process owner</td><td>Techniques
• Interview/workshop
• Data quality simulation</td></tr>
<tr><td colspan="3">Checklist
☐ Process activities and business transactions with known or presumed business problems (potentially caused by data defects) have been identified.
☐ Process metrics for monitoring the performance of the identified process activities have been identified.
☐ Business problems and data defects as potential cause have been described.
☐ Costs arising as a result of identified business problems have been described.</td></tr>
</table>

Table 5-2: Method Fragment for Activity I.1

A scope focuses on part of an organization's enterprise architecture (e.g. process activities, business objects, data object types) (Figure 5-1) in which business problems are known or presumed. The aim of applying the method is to identify causal chains (in particular business problems and the data defects that cause them) within these scopes and to specify DQ metrics for monitoring data defects.

> *Telephone Inc. has defined the availability check for IP-TV (process activity) as its scope. Potential customers can go onto a website and check whether IP-TV is available at a specific address (positive answer) or not (negative answer). A false negative answer, i.e. a negative answer despite IP-TV actually being available at the address in question, significantly reduces the probability of winning the inquirer as a new customer (value of a new customer approx. EUR 500 p.a.), and therefore makes it more difficult to achieve strategic corporate objectives (win new customers, increase revenues). The percentage of false negative answers is therefore a process metric for monitoring the performance of the availability check service. A clear correlation is presumed between the correctness of availability check answers and the quality (accuracy, completeness) of network infrastructure data.*

For each process activity (i.e. as part of every scope) focused on, at least one process metric should be identified for which measured values are available. Furthermore, the business problem (in the example, a false negative answer from the availability check) is to be described along with the impact of the problem on the process metric and the achievement of the corporate objective focused on.

> *In its search for critical data defects, Consumer Goods Inc. considered the entire supply chain and limited the scope by focusing on product data as well as intra-organizational and inter-organizational interfaces. The reasons for this definition of the scope were on the one hand the fundamentally good quality of the master data (little criticism, few known problems) and on the other, occasional but concrete criticism of the quality of logistical product data (in particular weights). Strategic corporate objectives (compliance, cost reduction, high service level) have not been refined by concrete process metrics.*

In addition to the focus on concrete process activities as the scope, it is also possible to concentrate on data object types (in the example of product data). This focus is useful if concrete data defects are known or if the DQM organization (e.g. department within the IT organization, shared service center structured according to data object types) calls for a focus on specific data objects.

> *Power Inc. has limited the scope to the business objects customer, contract and contract account as well as various business transactions from invoicing, customer service and marketing processes. The reason for this selection is the goal of being able to use the specified DQ metrics for comparing DQ measurements with other energy suppliers at a later date (benchmarking). The scope has therefore been limited to business objects and business transactions which are pre-*

> *sumed to be similarly structured by other energy suppliers and hence comparable.*

As a fundamental requirement, the overriding goal of DQ measurements should already be taken into account when defining the scope. While specifying DQ metrics as a DQM tool can be the only goal of the method application, in most cases concrete requirements such as improving the performance of a given process activity (cf. Telephone Inc.), concretizing known data defects and explaining the business relevance (cf. Consumer Goods Inc.) or inter-organizational benchmarking (cf. Power Inc.) are the drivers for the specification of DQ metrics.

5.2.1.1 Roles

The roles of the method are based on a reference model for data governance [Weber 2009, Weber et al. 2009]. The central role is the corporate data steward (cf. Figure 5-3). He or she is responsible for configuration (e.g. selection of techniques) and application of the method, coordinates cooperation between the people and boards involved, and documents the results. This role is usually assumed by the head of the DQM organization. In large organizations, however, this person may also take the Sponsor role and delegate the tasks of Corporate Data Steward to someone else in her or his team. Further description of the method assumes a DQM organization headed by the corporate data steward (one person) and a sponsor (one person, superior of the corporate data steward) outside of the DQM organization.

The sponsor defines goals for the method application, ensures the availability of the necessary resources and approves the business concept (result document of Activity III.2, cf. section 5.2.7). The prime task of the corporate data steward and sponsor in Activity I.1 is to identify a scope which on the one hand focuses on a major part of the company's value chain and for which on the other hand there are clear causal chains, i.e. clearly recognizable relationships between data defects and business-relevant problems.

> *The head of the DQM organization at Consumer Goods Inc. was in charge of applying the method with external support. As sponsor, the head of Supply Chain Management, which DQM also belongs to organizationally, supported the initiative (project budget, personnel resources, assistance with requests for company-wide interviews).*

Since, in addition to specific knowledge of the company's business processes, business model and organization structure, general experience with searching for causal chains is a success factor for application of the method, external assistance for the corporate data steward can be helpful.

Process owners [DIN/EN/ISO 2000] are useful interlocutors when identifying suitable scopes. Typical tasks of a process owner are the planning and monitoring of process goals, the maintenance of process metrics, and the planning and initiation of improvement initiatives. In terms of the method application, the process owner's task is to assess which process activities in particular could be jeopardized by data defects and which process metrics might clearly show the impacts of possible data defects.

5.2.1.2 Technique *interview/workshop*

The aim of an interview is to collect information on or evaluate a specific set of facts in a conversation with one or more people. A workshop, on the other hand, serves less for collecting information and more for developing or discussing outcomes (e.g. defining requirements). It is advisable to conduct a workshop with several people as discussions between the participants open up new perspectives and facilitate the identification of otherwise hidden correlations (e.g. interdepartmental causalities between data defects and business problems). The method provides interview guidelines, documentation templates and sample results (e.g. examples of causal chains) to assist with the envisaged interviews and workshops.

As neither interviews nor workshops are recommended exclusively in any of the activities involved, the method groups both options together in one *interview/workshop* technique and describes its application for each activity. To improve readability, either one interview or one workshop is described. In all cases, the corporate data steward conducts the workshop or interview, and tries to collect and document all the necessary information by means of targeted questions and discussions. Further techniques of the method (e.g. data quality simulation, cf. section 5.2.1.3) can assist an interview or workshop.

In order to identify possible scopes, an interview must first be conducted with the sponsor. This is used to develop

- a list of ongoing projects,
- a list of completed or failed projects, and
- a list of business problems

for which the data defects or DQM initiatives have already been discussed or are currently under discussion. The aim of this interview is to obtain knowledge of existing analyses, specifications or other project results and to identify process owners for further interviews. To assist the interview, the method provides examples of causal chains (see Appendix B.1) and a documentation template (see Appendix E.1). If concrete data defects or business problems are the reason for applying the method, this interview can be replaced by a brief agreement of the goals for application of the method.

In interviews with individual process owners, the previously sketched out scopes are to be concretized and for each scope

- a process activity,
- at least one business problem,
- at least one process metric whose measured values are impacted by the business problem,

- the impact of the business problems on process performance and
- data defects as the possible cause of each business problem

are to be described. The materials from B.1 and E.1 can be of assistance in this context.

5.2.1.3 Technique *data quality simulation*

This technique enables data quality metrics values to be simulated on the basis of a modeled DQM scenario. A scenario of this kind encompasses various causal chains (i.e. data defects, DQ metrics, business problems) and DQM initiatives, and defines assumptions in respect of the probability of data defects, the cause of business problems and the impact of DQM initiatives [Otto et al. 2010]. Figure 5-4 shows a tool for the modeling and simulation of DQM scenarios. The tool enables the definition of data defects and a probability can be given for each data defect (cf. slider in Figure 5-4). A scenario also encompasses business problems and the assignment of defined data defects as the cause of these business problems. A simulation step simulates the occurrence of data defects and the number of business problems caused. Reactive DQM initiatives can repair data defects and thus reduce the number of business problems caused in the next simulation step. The simulation of a concrete scenario for Telephone Inc. is described in section 5.2.7.

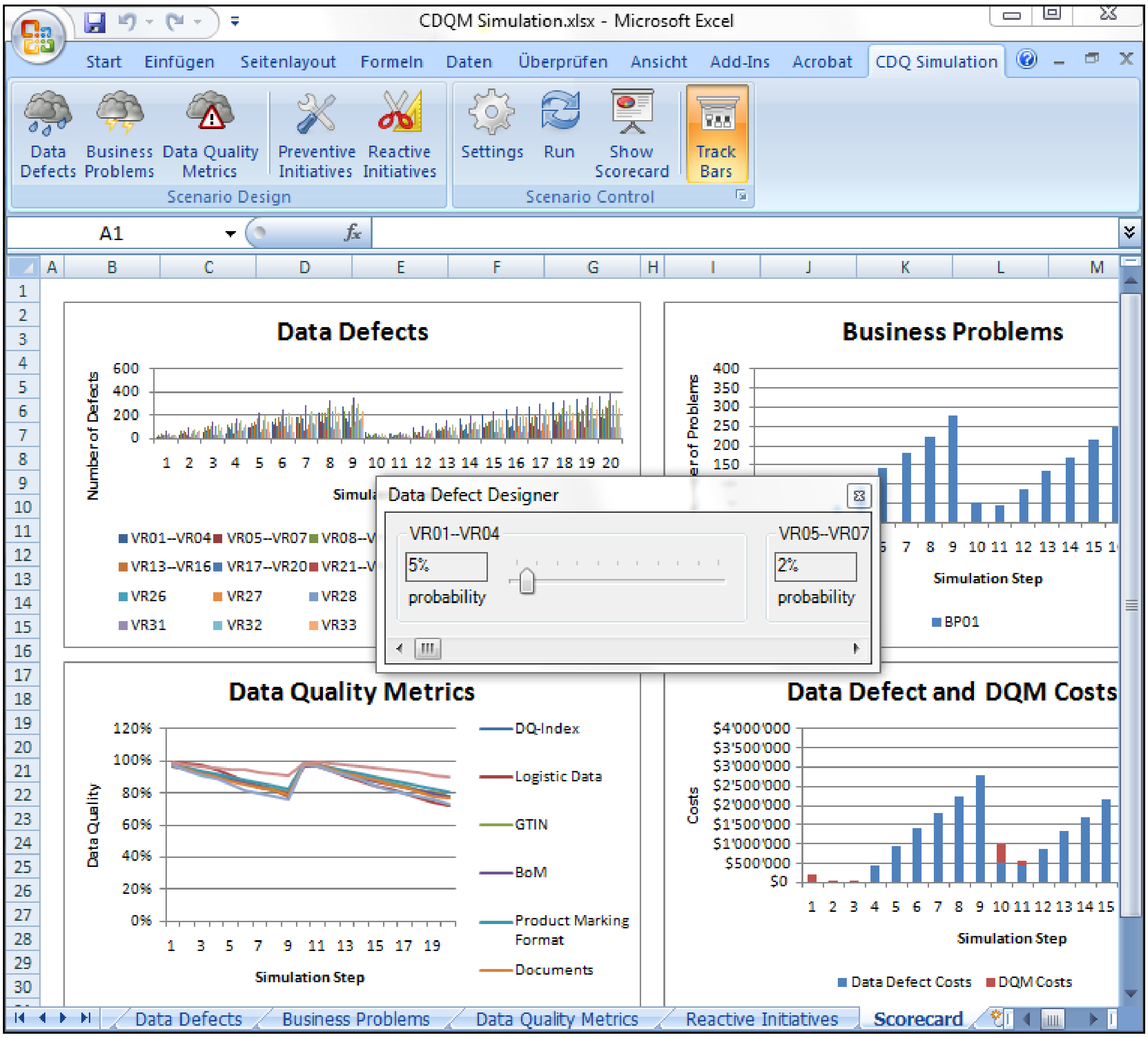


Figure 5-4: Tool for the Simulation of Data Quality Management Scenarios

When defining possible scopes, the tool shown in Figure 5-4 can be used to support the interview with sponsor and process owners, enabling discussed correlations to be illustrated and examples of causal chains explained. The method provides modeled scenarios for the causal chains described in Appendix B.1, which can be adapted and simulated with the tool (see Figure 5-4). A more detailed description of the technique and the simulation method used is given by OTTO ET AL. [2010].

5.2.2 Identify data and IT systems

The goal of Activity I.2 is to identify the IT systems used in the individual scopes and to obtain a description of the analyzed data which is as precise as possible (cf. Table 5-3). In interviews (cf. technique *interview/workshop*) with the respective process owners, the data used in a given process activity and the IT systems which deliver or use the data (databases or business applications respectively) are to be identified for each previously identified scope. In the case of data, not only the data object type (e.g. customer data), but also a detailed limitation of the presumably defective data objects (e.g. “all male customers in St. Gallen”, “all active products offered in Switzerland”) is to be documented. To assist the interview, the method provides a documentation template (see Appendix E.1).

<table>
<tr><td colspan="3">Activity I.2: Identify data and IT systems
The goal is to identify the data used in the process activities and the IT systems which store (databases) and/or use (business applications) the data for each scope previously identified, and to concretize these by limiting the data object types already documented. In addition, the type of possible data defects is to be limited by the selection of analyzed DQ dimensions.</td></tr>
<tr><td>Result document
List of scopes, data, IT systems and DQ dimensions</td><td>Roles
• Corporate data steward
• Process owner</td><td>Techniques
• Interview/workshop</td></tr>
<tr><td colspan="3">Checklist
☐ For each scope, the analyzed data have been described as precisely as possible (i.e. data object type and limitations).
☐ For each scope, the databases and business applications used have been identified.
☐ For each scope, the DQ dimensions to be focused on have been selected.</td></tr>
</table>

Table 5-3: Method Fragment for Activity I.2

> *Consumer Goods Inc. has selected approx. 100 data attributes (grouped into 20 data clusters) from the company’s product data model consisting of approx. 800 attributes. The strategy of the DQM organization is stated to center on “the timely provision of accurate, complete and consistent data”. In order to focus the discussion in planned workshops, the four DQ dimensions timeliness, accuracy, consistency and completeness have therefore been selected.*

In addition to limiting the data analyzed, the corporate data steward should select several DQ dimensions for the identification of causal chains in Activity I.3 in order to ensure a focused discussion. However, as the selection of approaches presented in section 4 already shows, no generally accepted definition of data quality exists. Instead, the various authors propose different lists of DQ dimensions which overlap in part. While the list of 15 dimensions proposed by WANG/STRONG [1996] is the most well-known approach, other works [Olson 2003, Price/Shanks 2005] should also be borne in mind, particularly taking into account DQ measurements (i.e. the operationalization of DQ dimensions through objectively measurable characteristics of data objects) as the goal of method application.

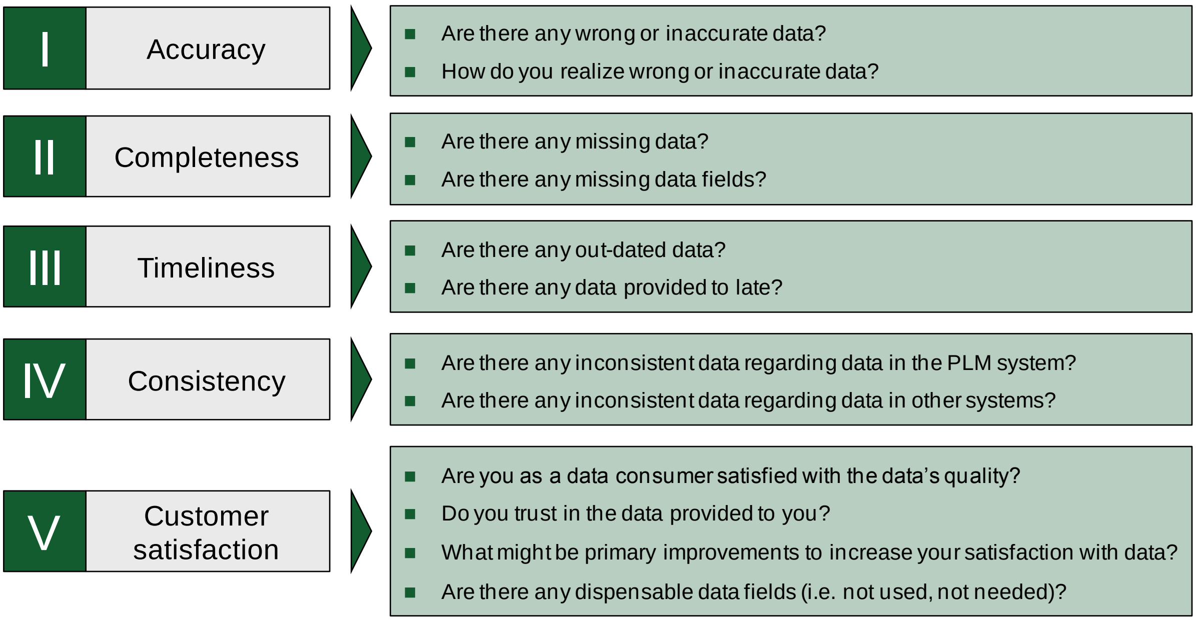


Figure 5-5: Guiding Questions for Selected DQ Dimensions

The method does not prescribe any selection of DQ dimensions, but provides various lists as a guide in Appendix A. The corporate data steward should select approximately four to eight dimensions from these suggestions and formulate guiding questions by way of explanation for each dimension (cf. Figure 5-5). The documentation template from Appendix E.2 can help in this context. If the DQM organization has formulated a DQM strategy, it is essential to take this into account as DQ dimensions (cf. Consumer Goods Inc.) it contains have already been aligned with corporate strategy as success factors for the creation of products and services.

5.2.3 Identify causal chains

The goal of this activity is to identify causal chains with business-critical data defects for which the corporate data steward will subsequently specify DQ metrics in Activity II.1 to monitor them (cf. Table 5-4). The identification process should be performed *top down*, i.e. the corporate data steward should try to identify data defects on the basis of known business problems.

<table>
<tr><td colspan="3">Activity I.3: Identify causal chains
The goal is to identify causal chains, i.e. cause-and-effect relationships between data defects, business problems and strategic corporate objectives. For this purpose, business problems are to be identified for each process activity within the scope, and for the each business problem, the impact (e.g. costs) and the data defects causing it.</td></tr>
<tr><td>Result document
List of causal chains</td><td>Roles
• Business data steward
• Corporate data steward
• Data user
• Technical data steward
• Process owner</td><td>Techniques
• Data quality analysis
• Data quality simulation
• Interview/workshop</td></tr>
<tr><td colspan="3">Checklist
☐ Causal chains have been described in full (i.e. data defects, business problems and impacts).
☐ Causalities (business problem / impact on strategic corporate objective, data defect / business problem) have been described precisely (i.e. quantified as far as possible).</td></tr>
</table>

Table 5-4: Method Fragment for Activity I.3

The technique *interview/workshop* supports the identification process through interviews with various experts within the company (cf. section 5.2.3.1). The technique *data quality simulation* can help the discussion in the interviews and illustrate the correlations discussed by means of simulated DQM scenarios. If the identification of data defects on the basis of business problems (top-down approach) does not lead to the desired result, or if the corporate data steward wants to identify potential data defects in preparation for interviews (bottom-up approach), the technique *data quality analysis* (cf. section 5.2.3.2) enables the identification of statistically striking data elements.

5.2.3.1 Roles

The identification of causal chains is the task of the corporate data steward. However, as in most cases he or she does not possess the required knowledge of the company's entire value chain with process and data architecture as well as the associated controlling systems (cf. section 5.1), this information has to be collected from different experts and consolidated.

Business data stewards are part of a company's DQM organization and report to the corporate data steward (cf. section 5.2.1.1). They are responsible for working with business departments and maintain, for example, business metadata or business data quality requirements to be met by application systems. Although other responsibilities are possible (e.g. per department, per process), further description of the method is based on the responsibility of a business data steward for one data object type in each case (e.g. material master, customer master).

A technical data steward represents the same for the IT department that a business data steward represents for the business department. Here again, further description of the method assumes one technical data steward per data object type. A technical data steward has detailed business knowledge of the databases and application systems which store or operate with the respective data.

Data users are company employees who create and use data in their day-to-day work (i.e. execute work procedures described in process activities), the emphasis being on use in the context of applying the method. As the employees are directly confronted with data defects in their work (e.g. missing, inaccurate, out-of-date data elements), they can describe the immediate impacts of defective data and are therefore important interlocutors when identifying causal chains.

> *In addition to the corporate data steward and external support, the project team at Consumer Goods Inc., who applied the method, includes a business data ste-*

> *ward for product data as well as a technical data steward. The technical data steward assisted with detailed expert knowledge of the functionality of the central product lifecycle management (PLM) system at Consumer Goods Inc. A total of 16 data users identified causal chains in seven workshops. Prior to the interviews, the sponsor communicated the necessity for the initiative and, in so doing, ensured the availability and active participation of the interviewees.*

5.2.3.2 Technique *data quality analysis*

The technique *data quality analysis* utilizes statistical methods to identify unusual data elements (i.e. which significantly deviate from the mean or are rarely used). Various software manufacturers offer applications to support this technique [Alur et al. 2007]. Providing that the corporate data steward considers its application useful, this technique should be used to prepare for a workshop (cf. section 5.2.3.3) in order to find points of reference for possible causal chains. In the case of identified deviations, however, it is important to check whether these really do involve a data defect and, if so, whether the data defect is attributable to a known business problem (bottom-up approach).

5.2.3.3 Application of the technique *interview/workshop* in Activity I.3

Knowledge of individual aspects is usually distributed among several people (technical data stewards have detailed knowledge of data defects; business data stewards and data users know about the business problems). Moreover, the processes often span different departments (e.g. address data entry by a call center, address data cleansing by the DQM organization and address data use by the invoice dispatch and marketing departments). As far as possible, therefore, the corporate data steward should try to bring together diverse roles and employees from different departments, and encourage discussion in a workshop.

> *Power Inc. conducted a workshop with a total of nine employees from the areas of business customer accounting, residential customer accounting, DQM, customer service, marketing, invoicing and invoice dispatch. For ten business transactions (e.g. mailing campaign, reminders, invoice compilation), the participants discussed business problems as well as identifying and evaluating defects in business partner data, contract data and contract account data.*

FERT, Basic Data	Material Number /NART	Material Description	Material Group	Material status	Material Type (MTART)	Sub Process Cluster
	Product Hierarchy	Authorization Group	ICP1/2	Cluster		

	Problem	Frequency rating	Impact rating		
			Costs	Service level	Complaints
Wrong data?	*description …*	(0 – 5)	(0 – 5)	(0 – 5)	(0 – 5)
Missing data?	*description …*	(0 – 5)	(0 – 5)	(0 – 5)	(0 – 5)
Out-dated data?	*description …*	(0 – 5)	(0 – 5)	(0 – 5)	(0 – 5)
Inconsistent data?	*description …*	(0 – 5)	(0 – 5)	(0 – 5)	(0 – 5)

Further questions ?

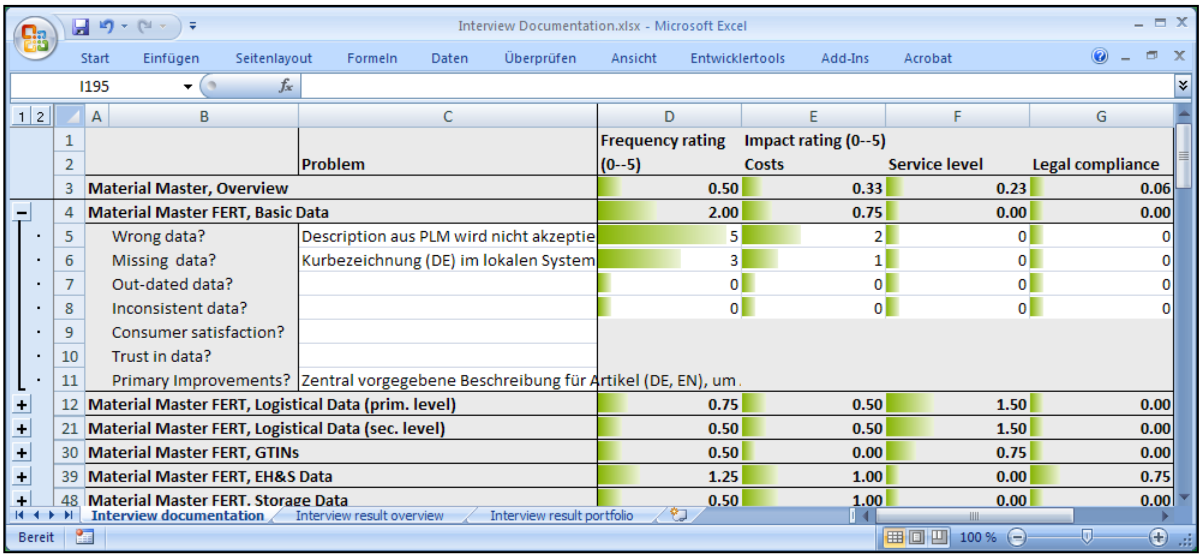
- Are you as a data consumer satisfied with the data's quality?
- Do you trust in the data provided to you?
- What might be primary improvements to increase your satisfaction with data?
- Are there any dispensable data fields (i.e. not used, not needed)?

Figure 5-6: Workshop Guideline for Identifying Causal Chains

To assist a workshop, the corporate data steward should set out the scope in a presentation (e.g. attributes, data object types and business transactions to be analyzed) so that he/she can focus the discussion. Consumer Goods Inc., for example, developed a guideline with attributes grouped according to business needs and then discussed this for each business transaction they were focusing on (see Figure 5-6).

> *Consumer Goods Inc. identified various causal chains in the seven workshops they conducted and also evaluated frequency and criticality (costs incurred directly by e.g. additional expense or faulty production, risk of complaints and impairment of service level). Figure 5-7 shows an extract from the documentation*

> *template used. The evaluation enabled Consumer Goods Inc. to identify seven problem areas which have structured the metrics system specified in Activity II.2.*



		Problem	Frequency rating (0--5)	Impact rating (0--5) Costs	Service level	Legal compliance
1						
2						
3	**Material Master, Overview**		**0.50**	**0.33**	**0.23**	**0.06**
4	**Material Master FERT, Basic Data**		**2.00**	**0.75**	**0.00**	**0.00**
5	Wrong data?	Description aus PLM wird nicht akzeptie	5	2	0	0
6	Missing data?	Kurbezeichnung (DE) im lokalen System	3	1	0	0
7	Out-dated data?		0	0	0	0
8	Inconsistent data?		0	0	0	0
9	Consumer satisfaction?					
10	Trust in data?					
11	Primary Improvements?	Zentral vorgegebene Beschreibung für Artikel (DE, EN), um				
12	**Material Master FERT, Logistical Data (prim. level)**		**0.75**	**0.50**	**1.50**	**0.00**
21	**Material Master FERT, Logistical Data (sec. level)**		**0.50**	**0.50**	**1.50**	**0.00**
30	**Material Master FERT, GTINs**		**0.50**	**0.00**	**0.75**	**0.00**
39	**Material Master FERT, EH&S Data**		**1.25**	**1.00**	**0.00**	**0.75**
48	**Material Master FERT, Storage Data**		**0.50**	**1.00**	**0.00**	**0.00**

Figure 5-7: Documentation Template for Evaluating Causal Chains

When evaluating causal chains, it is important to ensure equidistant rating levels and hence comparability of evaluation scales across the various workshops, for example by explaining how the rating levels in an employed Likert scale are to be interpreted. The corporate data steward should select rating dimensions and levels that are easy to understand and reflect corporate objectives (e.g. *frequency of business problems*, *impairment of customer satisfaction*, *product quality or service level*, *risk of complaints*, *extra work required*, *storage costs*, *manufacturing rejects* or *transport costs*). To ensure that evaluations are comparable, the corporate data steward should provide interpretation guidance, e.g. by mapping all rating dimensions onto costs (or another uniform criterion) and state the approximate costs for each rating level.

The type of technique applied will be largely dependent on the specific circumstances of the respective company (e.g. availability and distribution of employees), and therefore cannot be prescribed by the method. The following three application types have already been performed successfully (see Appendix E.5 for documentation templates).

- *Task-specific interview*. Interview with one or more company employees who perform a specific job (e.g. call center agent, customer service adviser). In interviews of this kind, the emphasis is not on discussing causal chains that span various departments or roles, but on concretizing known business problems and data defects. Appendix B.2 provides guiding questions for this purpose.
- *Department-specific workshop*. Workshop with several employees from a business department (e.g. finance, marketing, product development), a support function (e.g. decision-making support, quality management) or a production plant. It is useful to carry out this type of workshop when employees are geographically far apart, or when a large number of employees and business units are to be involved. By way of assistance, Appendix B.1 gives examples of causal chains for this purpose. In addition, the technique *data quality simulation* provides the possibility of simulating causal chains that are being discussed.
- *Interdepartmental workshop*. Workshop with several employees from different departments. Preference is to be given to this type of workshop as it allows the discussion to cover the entire scope. The corporate data steward should nonetheless limit the number of participants to between 8 and 15 people to ensure a focused discussion despite the welcome breadth.

Application of the method has shown that as well as the causal chains to be identified, possible initiatives for improving data quality are also discussed in a workshop. While this information is not directly necessary for the specification of DQ metrics, the corporate data steward should nonetheless document it in view of its usefulness for possible projects at a later date (cf. Figure 5-2 and documentation template in Appendix E.5).

5.2.4 Define and prioritize requirements

The goal of Activity II.1 is to define a list of requirements, i.e. a company-specific list of requirements to be met by a DQ metric, ranked in order of importance (cf. Table 5-6). After specifying the DQ metrics in Activity II.2, the corporate data steward checks that the requirements have been met in Activity III.1. Since these requirements represent what is expected of the DQ metrics, the corporate data steward should agree the list of requirements with the sponsor. Prioritizing the individual requirements also assists the approval and, if necessary, the adaptation of the list of requirements in Activity II.2.

<table>
<tr><td colspan="3">Activity II.1: Define and classify requirements
The goal is to define a list of requirements, i.e. a company-specific list of requirements to be met by a DQ metric, ranked in order of importance.</td></tr>
<tr><td>Result document
List of requirements</td><td>Roles
• Business data steward
• Corporate data steward
• Sponsor
• Technical data steward</td><td>Techniques
• Interview/workshop</td></tr>
<tr><td colspan="3">Checklist
☐ The defined requirements include all corporate guidelines (e.g. existing metrics systems, reporting templates, standards).
☐ The list of requirements does not contain any conflicting requirements.
☐ The prioritization reflects the importance of the individual requirements.</td></tr>
</table>

Table 5-5: Method Fragment for Activity II.1

> *Guidelines at Telephone Inc. require metrics to be integrated into a Balanced Scorecard used throughout the company. This requirement implies the necessity for taking into account the structure and units of measure of the Balanced Scorecard metrics system when specifying DQ metrics (particularly when selecting the level of measurement and unit of measure). For this reason, Telephone Inc. has defined the use of a percentage scale (optimum: 100%) as a requirement for the design of the DQ metrics.*

Typical requirements relate to the presentation of measured values, subsequent areas of application or the structure of the metrics system. To support the activity, the method includes a list of requirements to be met by DQ metrics (see Appendix C). These requirements can assist the discussion in a workshop (cf. technique *interview/workshop*, section 5.2.1.2) between the corporate data steward and the various other data stewards (those responsible for the data object types affected by the causal chains). Appendix E.4 provides a template for documenting the requirements.

> *Power Inc. formulated the comparison with other companies (benchmarking) as the goal of using DQ metrics. To ensure the subsequent comparability of DQ measurements, validation rules should only consider attributes of standardized data models (e.g. data fields of unmodified standard software).*

The specification of DQ metrics in the form proposed by the method, i.e. with the identification of causal chains (cf. sections 5.2.1 and 5.2.3), the use of validation rules as measurement technique (cf. section 5.2.5.1) and the mapping of metrics values onto a percentage scale (cf. 5.2.5.2), fundamentally ensures the fulfillment of various requirements.

- *Aggregation*. The use of a uniform calculation formula (cf. technique *specification of calculation formulas*, section 5.2.5.2) in the measuring technique of DQ metrics fundamentally ensures that it is possible to aggregate the measured values. However, it may be necessary to bear in mind different limitations of the validated data (cf. Activity I.2).
- *Acceptance.* The discussion on possible DQ measurements prompted by the interviews or workshops in Activity I.3 facilitates introduction of the metrics and promotes the acceptance of their use.
- *Improvability*. As the fundamental structure of the DQ metrics is based on data defects, the result of the specification process provides concrete points of reference for DQM measures to eliminate or prevent the monitored defects for each metric.

- *Business relevance*. The identification of business-critical data defects (cf. Activity I.3) and the documentation of causal chains (cf. Activity II.2) ensure that each validation rule and each DQ metric can be assigned to at least one business problem.
- *Measurability*. The specification of validation rules as measuring techniques (cf. technique *specification of validation rules*, section 5.2.5.1) ensures that values can be determined for the specified DQ metrics.
- *Standardization*. The use of the calculation formulas explained in section 5.2.5.2 ensures the interpretation of DQ measurements on a ratio scale and thus the comparability of metrics values.

5.2.5 Specify data quality metrics

The goal of Activity II.2 is to specify DQ metrics for monitoring the previously identified data defects (cf. Table 5-6). For this purpose, it is necessary to describe the databases for whose data elements the metrics values are to be measured (measuring points), the data to be validated, the timing of the measurements (measuring frequency), a measuring scale, a unit of measure and a measuring technique.

<table>
<tr><td colspan="3">Activity II.2
The goal is to specify at least one DQ metric for each previously identified causal chain. For this purpose, it is necessary to describe the databases for whose data elements the metrics values are to be measured (measuring points), the data to be validated, the timing of the measurements (measuring frequency), a measuring scale, a unit of measure and a measuring technique. In terms of measuring techniques, the method proposes validation rules and supports the design of rules and calculation formulas by describing two techniques.</td></tr>
<tr><td>Result document
List of DQ metrics specifications</td><td>Roles
• Business data steward
• Corporate data steward
• Technical data steward</td><td>Techniques
• Data quality simulation
• Interview/workshop
• Specification of calculation formulas
• Specification of validation rules</td></tr>
<tr><td colspan="3">Checklist
☐ Each DQ metric has been specified in full, i.e. the data to be validated, a measuring point, a measuring scale, a unit of measure, a measuring frequency and a measuring technique have all been described.
☐ The specification of the measuring techniques takes into account all previously documented causal chains, unless reasons are stated as to why a causal chain has not been included.
☐ All specified measuring techniques can be implemented at the specified measuring points.</td></tr>
</table>

Table 5-6: Method Fragment for Activity II.2.

A prime task of the corporate data steward is the specification of measuring techniques, i.e. the design of procedures that investigate whether or not the examined data show one or more of the business-critical data defects identified in Activity I.3. While the attributes *limitations*, *measuring frequency* and *measuring point* stated in the meta-model (cf. Figure 5-2) are also decisive for implementing the metrics later on, it is often the case that they can only be concretized during the course of initial tests (e.g. validating individual data objects with specified validation rules). Two types of measuring technique can be distinguished:

- *Data user surveys*. The quality of data can be evaluated by the people who use it in interviews or surveys [Lee et al. 2002, Price et al. 2008]. An interview guideline or a questionnaire is therefore a possible measuring technique that is to be selected in particular when the identified data defects have to be assigned to DQ dimensions such as *objectivity*, *believability* or *ease of understanding* which are (*subjectively* according to PRICE ET AL. [2005]) difficult to interpret in technical terms.
- *Validation by means of rules*. The quality of a data object can be measured by examining specific characteristics of the data elements or attributes [Loshin 2001, Olson 2003, Fan et al. 2008]. If a data object shows all the required characteristics, its quality is 100%. Data characteristics (e.g. last time a data element was modified, completeness of a data element) can be formulated as rules and validated automatically. Validation rules of this kind for verifying the characteristics required for a data object (in a given context) are therefore a possible measuring technique.

Both techniques are fundamentally suitable for checking data for possible defects. However, since interviews or surveys tie up resources for each measurement (e.g. for the consultation) and consequently cannot be conducted frequently, the presented method focuses on using validation rules as the preferred technique for measuring data quality. Several authors [Lee et al. 2002, Price et al. 2008] provide guiding questions as preparation for interviews and surveys.

The technique *specification of validation rules* supports the corporate data steward with the formulation of validation rules. If the information collected in Activity I.3 is not sufficient, he/she can also use the Technique interview/*workshop* e.g. to conduct interviews with data stewards in order to limit the data to be validated more precisely, to check the availability of data elements and metadata (e.g. time when an attribute was last changed) at the possible measuring points, or to obtain data models for exact designations of attributes. As the way of applying the technique *interview/workshop* is heavily dependent on the result of Activity I.3, the method does not provide further assistance here in the form of guiding questions or specific documentation templates. The corporate data steward can nonetheless use the documentation templates in Appendix E.5 as a guide and check what information is already available and what is not.

In order to be able to make statements on data quality with business relevance (e.g. "the quality of data required for letters to customers is only 70%"), the corporate data steward should bundle specific validation rules and consolidate the verification of individual rules into one metric value (cf. technique *specification of calculation formulas*, section 5.2.5.2).

> *Analysis of the identified causal chains at Consumer Goods Inc. resulted in 32 validation rules for 7 problem areas (e.g. dangerous goods indicators, quality of bills of materials, customs information). For each problem area, a calculation formula (same calculation pattern for all metrics) aggregates the results obtained with the assigned validation rules into one metric value.*

To prevent the metrics system from becoming too complex (i.e. not adding still more calculation layers), it is useful to formulate exactly one calculation formula per data quality metric and to use the same pattern for all calculation formulas (cf. section 5.2.5.2). This will also mean that all metrics use the same scale and unit of measure. Appendix E.5 provides documentation templates for specifying validation rules and aggregating the rules into one calculation formula to help the corporate data steward.

5.2.5.1 Technique *specification of validation rules*

This technique assists the corporate data steward with the specification of validation rules as measuring technique for DQ metrics and uses the result documents produced in Activity I.3. The goal is to specify one or more validation rules for each business-critical data defect identified in Activity I.3.

```
for each data object d in database

    // Beispiele für Wertprüfungen
    if d.Attribute == NULL         : return 0
    if d.Attribute  < 42           : return 0
    if d.Attribute != 'foo'        : return 0
    if !List.Contains(d.Attribute) : return 0

    // Beispiele für Wertvergleiche
    if d.Attrubute1      != d.Attribute2 : return 0
    if d.Attribute1       < d.Attribute2 : return 0
    if d.Attribute1 + 42 != d.Attribute2 : return 0

    // Beispiele für metadataprüfung and -vergleich
    if d.Attribute.Type != 'numeric'          : return 0
    if d.Attribute.LastChange > Now - 2 weeks : return 0
    if d.Attribute.NumberOfChages > 42        : return 0
    if d.Attribute1.Type != d.Attribute2.Type : return 0

    return 1 // data object ist korrekt

end for
```

Figure 5-8: Validation Rule Patterns with Examples

A data defect is a characteristic of an attribute or data element which does not meet the (context-specific) requirements specified for the associated data object type. A missing data element of an attribute for representing a customer’s first name, for example, is a data defect if customers are to receive personally addressed letters in a process activity. When compiling an invoice, the first name is not necessary (unlike when sending out an invoice) and therefore the absence of this data element is not a data defect.

Validation rules can uncover data defects by successively evaluating the data objects (e.g. customer data). In concrete terms, a validation rule examines whether an attribute or data element possesses a specific characteristic. The rules therefore serve the purpose of validating data – if no representative data defect is found in a data object by means of the validation rules, the data object is considered to be correct (cf. differentiation in section 5.2.5.2). Based on literature research (cf. section 3) and project experience (cf. examples), the following validation rule patterns can be distinguished. Figure 5-9 shows examples of rules for the three validation rule patterns in a pseudocode notation.

- *Data element validation*. The rule checks whether a data element is present or whether a data element lies within the required domain (e.g. set of predefined data elements, range of numerical values).
- *Value comparison*. The rule checks whether the data elements of two or more attributes are equal, or whether the distance between them lies within a specific tolerance.
- *Metadata validation and comparison*. The rule determines the metadata of an attribute and performs a data element validation or a value comparison for this metadata.

Appendix D contains various examples of validation rules and can help the corporate data steward with specification. Appendix E.5 also provides a documentation template which contains a simple pattern for documenting attributes to be validated and compared (cf. Figure 5-9).

The corporate data steward can use this template in interviews (cf. section 5.2.1.2) with the data stewards who are responsible for the data object type being focused on so that discussed data defects can be directly documented in a structure resembling the verification and comparison logic of the validation rules. Another documentation template from Appendix E.5 supports the assignment of validation rules to DQ metrics, which has already been implicitly prescribed by the causal chains identified in Activity I.3, but should nonetheless still be explicitly documented after specifying the validation rules.

Data Quality Metric	Accuracy	Consistency	Completeness	Number of Rules
Business Partner	VR03, VR06	VR07	VR01, VR02, VR04, VR05	7 Rules
Contract	VR08	—	—	1 Rules
Contract Account	VR10	VR09, VR11	—	3 Rules
New Customer (move in)	VR03, VR06	VR07	VR01, VR02, VR04, VR05	7 Rules
Accounting	VR08, VR10	VR07	—	3 Rules
Postal Marketing Campaign	VR03, VR06	—	VR01, VR02, VR04, VR05	6 Rules
Payment	—	VR11	—	1 Rules
Maintenance (Customer Information)	VR03, VR06	VR07	VR01, VR02, VR04, VR05	7 Rules
Invoicing	VR03, VR06	VR09	VR01, VR02, VR04, VR05	7 Rules
Customer Magazine Delivery	VR03, VR06	VR07	VR01, VR02, VR04, VR05	7 Rules
DQ Index	VR03, VR06, VR08, VR10	VR07, VR09, VR11	VR01, VR02, VR04, VR05	11 Rules

Legend:
VR01: Salutation role person (completeness)
VR02: Salutation role organization (completeness)
VR03: Salutation (accuracy)
VR04: Name (completeness)
VR05: Address (completeness)
VR06: Address (accuracy, duplicates)
VR07: Language (consistency)
VR08: Reason for accounting blocking (accuracy)
VR09: Invoice / dunning recipient (consistency)
VR10: Invoice form (accuracy)
VR11: Payment method (in and out, consistency)

Figure 5-9: Assignment of Validation Rules (VR) to DQ Metrics

5.2.5.2 Technique *specification of calculation formulas*

When data objects are verified by means of validation rules, a calculation formula links the results of individual validation rules to a metric value. As already mentioned above, it is advisable to use precisely one calculation formula per metric. In this case, the calculation formula of the DQ metric calculates the value of the metric on the basis of the results of validation rules which are assigned to that metric.

As a fundamental principle, the calculation formula should be readily comprehensible in order to secure the acceptance of newly specified DQ metrics. Depending on subsequent applications (e.g. use in target agreements), however, metrics values should also make the impacts of performed initiatives visible and comparable over long periods of time. To ensure that the metrics are readily understood, clear statements should be possible such as “a data object is correct if it has no known defects”.

The following proposals for calculation formulas assume that a data quality metric m is assigned to a specific number of validation rules $r \in R|_m$, whereby the individual subsets $R|_{m_1}, R|_{m_2}, \ldots, R|_{m_n}$ of the set of all rules R must not be disjoint (multiple assignments can be useful: Complete address data, for example, may be required both for sending invoices (metric 1) and for specific marketing initiatives (metric 2)). Furthermore, the following validation rule pattern is assumed (cf. Figure 5-8).

$$r(d) = \begin{cases} 0, \text{ if } d \text{ has the defect represented by } r. \\ 1, \text{ if } d \text{ has not the defect represented by } r. \end{cases} \tag{4–1}$$

A simple possibility for calculating a value of the metric m to measure the quality of the data objects D (e.g. a company’s customer data) is the ratio of the number of defective data objects $\left|D|_{\text{defect}}\right|$ to the total number of data objects $|D|$. If this value is subtracted from 1, the proportion of correct data objects is obtained, i.e. a value that can be interpreted as *data quality* $m(D)$ with $0 \le m(D) \le 1$ on a ratio scale. If, for example, 58 of 100 validated customer data objects are defective, the statement “the data quality of customer data is 42%” would be possible with this calculation formula.

$$m(D) = 1 - \frac{\left|D|_{\text{defect}}\right|}{|D|} \tag{4–2}$$

With this simple calculation formula, it is decisive to define the circumstances under which a data object is correct. A simple and readily understandable definition is that a data object is correct from the viewpoint of a DQ metric m precisely if it has no known data defect, i.e. if no validation rule $r \in R|_m$ finds a defect when validating the data object d.

$$D|_{\text{correct}} = \left\{ d \in D \mid r(d) = 1 \; \forall r \in R|_m \right\} \tag{4–3}$$

> *Power Inc. decided that a data object should only be defined as correct if it has no known data defect. This decision was motivated in particular by the clarity of the definition and the fact that it was therefore simple to communicate. In order to be able to nonetheless prove impacts of DQM initiatives that are not visible in DQ measurements due to this restrictive definition, metrics values of validation rules, and not just the aggregated values of metrics, should be stored and visualized.*

The risk of the outlined restrictive definition is that impacts of DQM initiatives (e.g. the elimination of duplicates) will not become visible in the sequence of values of a DQ metric since defects that are yet to be eliminated will hide the effect. A threshold value s for the proportion of tolerated data defects in the definition of correct data objects can increase the sensitivity of the DQ metric.

$$D|_{\text{correct}} = \left\{ d \in D \;\middle|\; \frac{\sum_{i=1}^{|R|_m|} r_i(d)}{|R|_m|} \leq s \right\} \tag{4–4}$$

Another problem with the calculation formulas described thus far is the fact that all validation rules assigned to a DQ metric m are verified for all data objects of the data object type D. While a general limitation of the data object type is possible (e.g. *only active data objects* or *only data objects of products sold in Europe*), it is not possible to differentiate the data objects to be validated (i.e. assign specific data objects $d_j \in D|_{r_i}$ to specific validation rules

$r_i \in R|_m$). The following calculation formula contains a differentiation of this kind, but at the expense of a more complex calculation formula that is consequently more difficult to communicate.

$$m(D) = 1 - \frac{\sum_{i=1}^{|R|_m|} \left(\frac{\sum_{j=1}^{|D|_{r_i}|} r_i(d_j)}{|D|_{r_i}|} \right)}{|R|_m|} \qquad (4\text{–}5)$$

> *In the case of Consumer Goods Inc., the identified data defects mostly relate to very specific problems which can only occur in a small number of data objects. Thus, for example, missing dangerous goods indicators were identified as a business problem, but only for small series produced especially for marketing purposes. Customs data (e.g. country of production) that is not available on time also leads to business problems, but only in the case of data objects relating to products that are transported across certain national borders. However, as verification of the specified validation rules for all data objects would lead to very low error rates (>99% data quality), Consumer Goods Inc. decided to specify the data objects to be validated as precisely as possible for each validation rule.*

Weighting factors $\omega_i \geq 0$ also enable the influence of individual rules r_i on the calculated metric value to be controlled (e.g. a rule for validating dangerous goods indicators could be given a higher weighting than a rule for validating the correct product classification). At the same time, however, it is important that all weighting factors should always be compared with each other and the intended impacts (e.g. "r_1 is twice as important as r_2") documented for subsequent changes.

$$m(D)=1-\frac{\sum_{i=1}^{|R|_m}\left(\frac{\sum_{j=1}^{|D|_{r_i}}\omega_i r_i(d_j)}{|D|_{r_i}}\right)}{\sum_{i=1}^{|R|_m}\omega_i} \tag{4–6}$$

5.2.5.3 Application of the technique *data quality simulation* in Activity II.2

The simulation of various calculation options, particularly the selection of weighting factors and threshold values, can help with the selection and configuration of the calculation formula.

> *Power Inc. simulated DQ measurements with different calculation formulas to illustrate the configuration of the selected formula. The result of validating 100 data objects using 11 validation rules and the values of 10 DQ metrics calculated on this basis were simulated. The employed calculation formula is based on the formulas (4–2) and (4–4). Figure 5-10 shows simulated values for two metrics, with a threshold value of 0% (green) and 35% (blue) in each case.*

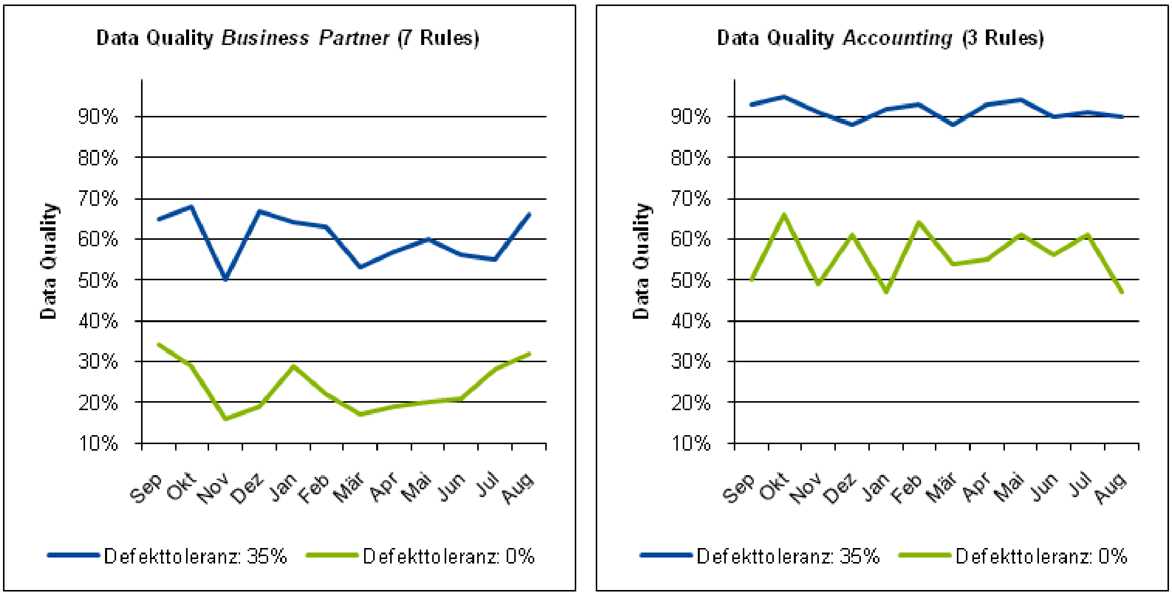


Figure 5-10: DQ Measurement with Different Calculation Formulas

5.2.6 Check the fulfillment of requirements

Activity III.1 checks whether the requirements defined in Activity II.1 to be met by the DQ metrics specified in Activity II.2 are fulfilled (cf. Table 5-7). If high-priority requirements are not fulfilled, the requirements, their prioritization and the metrics specifications have to be reviewed and adapted (i.e. return to Activity II.1). As outlined in section 5.2.4 for Activity II.1, the corporate data steward should hold a workshop and jointly check with the data stewards whether requirements are fulfilled (cf. section 5.2.1.2). The documentation template from Appendix E.6 supports this activity.

<table>
<tr><td colspan="3">Activity III.1: Check the fulfillment of requirements
The goal is to check whether the requirements defined in Activity II.1 are fulfilled. If high-priority requirements are not fulfilled, the requirements, their prioritization and the metrics specifications have to be reviewed and adapted (i.e. return to Activity II.1).</td></tr>
<tr><td>Result document
List of evaluated requirements</td><td>Roles
• Business data steward
• Corporate data steward
• Technical data steward</td><td>Techniques
• Interview/workshop</td></tr>
<tr><td colspan="3">Checklist
☐ All high-priority requirements have been fulfilled.
☐ For all unfulfilled requirements, a reason has been stated for each metric that does not fulfill the requirement.</td></tr>
</table>

Table 5-7: Method Fragment for Activity III.1

5.2.7 Document specification

Activity III.2 is the last activity of the procedure model and therefore delivers the result of applying the method. The goal of this activity is to document the specifications of the DQ metrics as an implementable business concept, i.e. in the form of a document on the basis of which the specified measuring techniques and scales can be implemented (cf. Table 5-8).

<table>
<tr><td colspan="3">Activity III.2: Document specification
The goal is to document the specifications of the DQ metrics as an implementable business concept, i.e. in the form of a document on the basis of which the specified measuring techniques and scales can be implemented.</td></tr>
<tr><td>Result document
Business concept for implementing DQ metrics</td><td>Roles
• Sponsor
• Corporate data steward</td><td>Techniques
• Documentation
• Data quality simulation</td></tr>
<tr><td colspan="3">Checklist
☐ All specified DQ metrics have been documented, i.e. in particular measuring techniques and scale for each metric.
☐ A causal chain has been documented for each DQ metric, i.e. the impact of the data defects it monitors.</td></tr>
</table>

Table 5-8: Method Fragment for Activity III.2

Activity III.2 (and with it the entire method) concludes with the release of the documented metrics specification (if necessary, after adjustments) by the sponsor.

5.2.7.1 Technique *documentation*

As companies usually use templates that have been agreed with their IT department or external IT providers for business concepts, Appendix E.7 only contains a generic structure for a business concept with notes on using the result documents produced during the course of applying the method.

Using the technique *data quality simulation* makes it possible to illustrate the elements identified and specified (causal chains, validation rules, calculation formulas, etc.). Modeling various scenarios and causal chains is useful, particularly for consultation with the sponsor.

> *Telephone Inc. identified six business problems (errors in the availability check) and specified nine validation rules as the measuring technique for a DQ metric. On the basis of error reports it is possible to estimate the probability of individual defects occurring. Figure 5-11, Figure* 5-12 *and Figure 5-13 show a simulation of the data defects, the business problems caused as a result and the sequence of values of the DQ metric. Pairwise disjoint validation rules are formulated for monitoring the data defects. This means that if a data defect occurs, it is bound to cause an error in the availability check and no inconsistency is monitored by more than one data defect.*

	Defect 1	**Defect 2**	**Defect 3**	**Defect 4**	**Defect 5**	**Defect 6**	**Defect 7**	**Defect 8**	**Defect 9**
Risk	953, 61	734, 50	1146, 127	1028, 37	428, 35	643, 71	495, 84	761, 129	369, 62
Effect	Error 1	Error 1	Error 2	Error 3	Error 3	Error 4	Error 5	Error 5	Error 6

Table 5-9: Parameters (Mean, Variance) of Normally Distributed Defect Frequencies

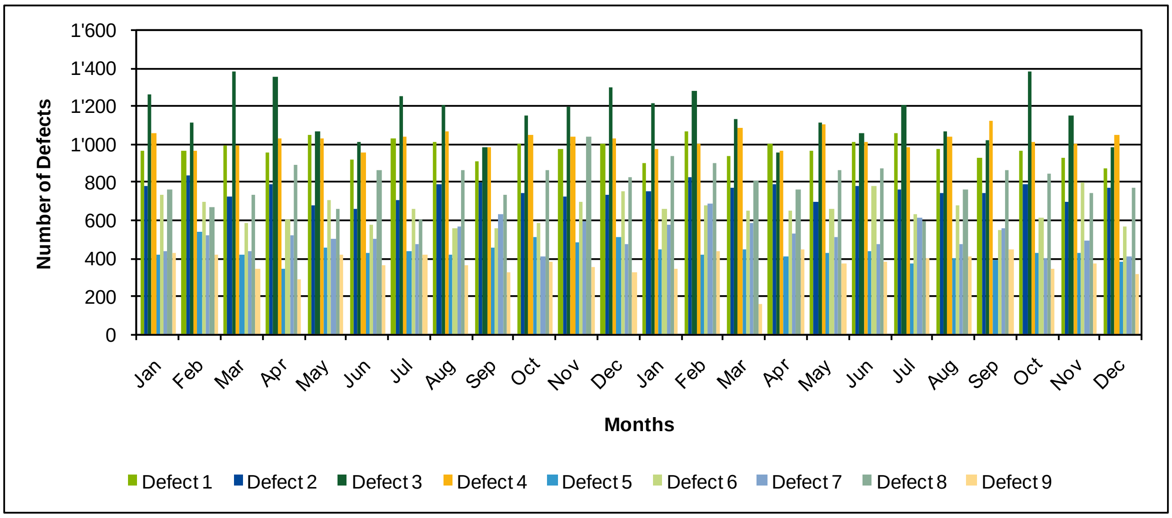


Figure 5-11: Simulation of Data Defects

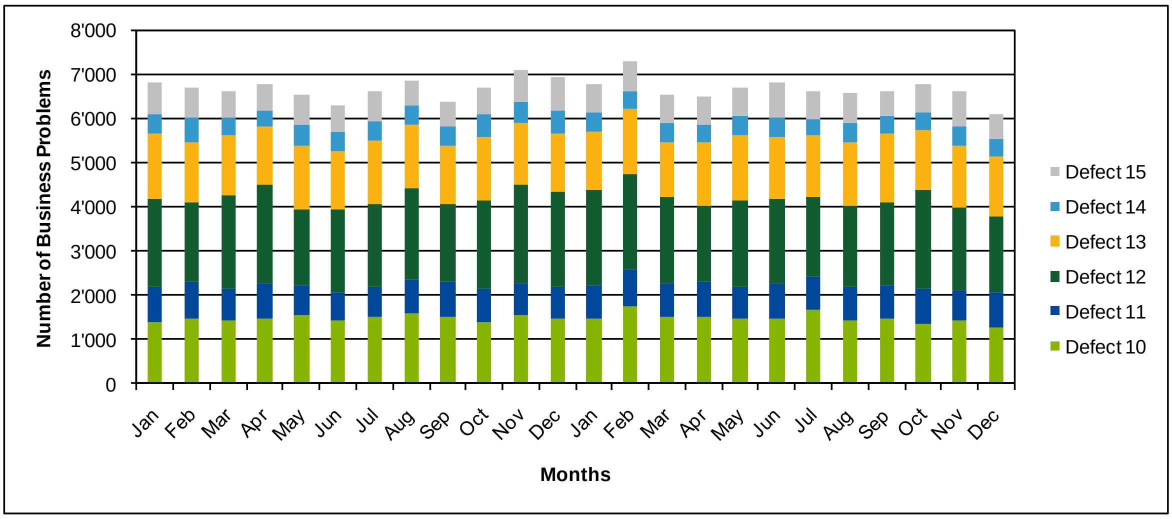


Figure 5-12: Business Problems Caused by Data Defects

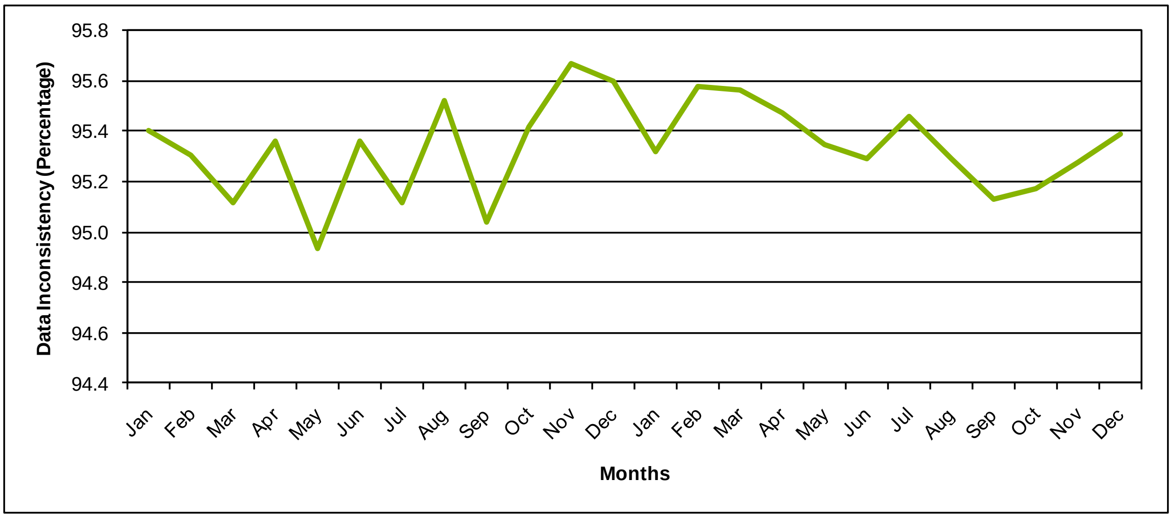


Figure 5-13: Monitoring of Data Defects with a DQ Metric

6 Method Evaluation

On February 9, 2010, eleven business experts from ten companies discussed the method and evaluated its benefits in a 90-minute focus group interview as part of a two-day workshop conducted by the CC CDQ. Appendix F gives the participants (see Table F-1) and the guiding questions (see Table F-2) for the focus group interviews. The guiding questions for evaluating the business value are based on the guidelines for orderly modeling [Schütte/Rotthowe 1998] and evaluation criteria for reference models [Frank 2007].

All the business experts confirmed the necessity for DQ measurements and also for methodological support when designing company-specific DQ metrics (cf. Table 6-1). The alignment of DQ metrics with strategic corporate objectives was also confirmed as a major design goal.

Relevance Perspective
• DQ measurements are important in order to replace reactive DQM by the preventive identification of critical data defects. • The method helps to reduce the complexity of specifying DQ metrics for monitoring business-critical data defects. • With the specification of business-oriented DQ metrics, the method addresses an important task, but does not adequately consider dynamic process changes which cause the identified causal chains to “age”. • The consideration of causal chains between data defects and the performance of business processes is a success factor for the implementation of effective DQ metrics.

Table 6-1: Evaluations of Business Experts for the Relevance Perspective

In view of the high relevance of DQ measurements, the requirement for a meticulous cost-benefit analysis for the specification and implementation of DQ metrics was ranked as low by some business experts (cf. Table 6-2). During the discussion, interviewees from companies which operate in heavily regulated markets (e.g. the pharmaceutical and chemical industries) in particular described data defects that can lead directly to infringements of legal requirements (e.g. the declaration of harmful substance content in plant protection products or shelf-life information). For the specification of DQ metrics for monitoring data defects of this kind, the business experts consider a cost-benefit analysis to be superfluous as monitoring is a legal requirement for the companies concerned. Other business experts confirmed the necessity for cost-benefit analyses and pointed out the possibility of embedding the specification of DQ metrics in other projects. A company in the chemical industry, for example, specified validation rules during the course of purchasing process harmonization [Ebner et al. 2011].

Economic Perspective
• A cost-benefit analysis is only necessary for less critical data defects. Compliance with legal requirements, for example, must be monitored. • Validation rules for monitoring data defects can be a by-product of completed or planned projects (e.g. harmonization of business processes, implementation of a new ERP system). The identification of such projects can reduce the anticipated effort required for the specification and implementation of DQ metrics.

Table 6-2: Evaluations of Business Experts for the Economic Perspective

The complexity of the method was ranked as critical by some business experts as they feared a high requirement for explanations when carrying out the individual activities (cf. Table 6-3). However, no superfluous or redundant elements were identified in a discussion on complexity. Other business experts, on the other hand, took a positive view of the communication required between different roles within the company when applying the method as this raises the fundamental awareness for data quality.

Deployment Perspective
• Monitoring the work performed by individual employees, i.e. the data quality resulting from data input and data maintenance processes, is always a delicate issue. In order to win the support of the employees involved, the goal of DQ measurements needs to be explained and illustrated in a readily understandable manner in interviews and workshops. • The process of identifying business-critical data defects raises the awareness for and the attention to data quality, and helps to explain the importance of data quality for the company's success. • The method would be more comprehensible with fewer components and more examples. • In order to raise problem awareness in the context of data defects, the results of DQ measurements should be reported to the employees who input data.

Table 6-3: Evaluations of Business Experts for the Deployment Perspective

While the engineering perspective was also a controversial issue in the discussion, the business experts were unable to identify any concrete requirements for modification (cf. Table 6-4). A data customer/supplier perspective was proposed as a concrete addition to support the identification of causal chains. Representatives of both sides should discuss the causes and effects of data defects in workshops to identify causal chains so that preventive initiatives as well as causal chains can be identified. Furthermore, some business experts suggested that validation rules for general requirements (e.g. for monitoring legal compliance) should be provided in order to speed up the specification process.

Engineering Perspective
• Business problems are a good starting point for identifying critical data defects. • There are too many roles; four roles should be sufficient. • A collection of general DQ metrics, e.g. for monitoring compliance with legal requirements, would be desirable. • The roles should include a data customer and a data supplier perspective. • The method should also include the definition of reporting channels for the results of DQ measurements. • Monocausal relationships between one business problem and one data defect (which only relates to one DQ dimension) might not be sufficient for modeling real problems. Business-critical causal chains can be more complex.

Table 6-4: Evaluations of Business Experts for the Engineering Perspective

The business experts did not discuss the epistemological perspective of the method (cf. Appendix F, Table F-2) as the quality of the design process did not affect their overall evaluation of the method. In the view of the focus group, further examples of DQ metrics and their implementation would more useful than the explanation of design decisions.

Overall, the focus group confirmed the need for business-oriented DQ metrics and for methodological support for their specification. While the method itself was seen as complex and requiring explanation, it was nonetheless considered to be fit for purpose and useful. A need for extension was identified in particular with regard to collecting examples of causal chains (cf. Appendix B.1) and validation rules (cf. Appendix D).

7 Conclusions and Further Research Requirements

This paper describes a method for identifying business-critical data defects and for specifying DQ metrics for monitoring the identified data defects. As measuring technique, the method proposes validation rules which can be used by a measuring system to verify characteristics of data elements and to identify data defects. The DQ metrics specified by the method application enable an objective evaluation of the quality of selected data elements and consequently comparable DQ measurements.

The method is a complex artifact whose application involves effort. In individual cases, therefore, it is important to weigh up which method elements actually produce business benefits. Thus, for example, it is not absolutely necessary to create all result documents or to define requirements for data quality metrics and verify their fulfillment with each application of the method. The method elements are instead to be seen as a structured collection of instructions and techniques which the corporate data steward can adapt in line with company-specific requirements. At the same time, however, the method sets forth a meaningful structure for the elements which should only be changed for good reason.

Further research is required in particular with regard to extending the collection of causal chains and validation rules. Application of the method in more cases will deliver the necessary results for this purpose as well as for evaluating the method itself. The specification of validation rules for general (i.e. not company-specific) data quality requirements (e.g. for legal requirements, cf. section 1) will enable the same DQ metrics to be used in different companies. This will make it possible to compare the DQ measurements of different companies and to identify successful DQM structures and initiatives. Moreover, the effort required for the specification of DQ metrics for individual companies will be reduced.

Literature

[Alur et al. 2007]
Alur, N., Joseph, R., Mehta, H., Nielsen, J. o. r. T., Vasconcelos, D., IBM WebSphere Information Analyzer and Data Quality Assessment, IBM, 2007
[Ballou et al. 1998]
Ballou, D., Wang, R. Y., Pazer, H., Kumar.Tayi, G., Modeling Information Manufacturing Systems to Determine Information Product Quality, in: Management Science, 44, 1998, No. 4, pp. 462-484
[Baskerville 1997]
Baskerville, R. L., Distinguishing action research from participative case studies, in: Journal of Systems and Information Technology, 1, 1997, No. 1, pp. 24-43
[Batini et al. 2007]
Batini, C., Barone, D., Mastrella, M., Maurino, A., Ruffini, C., A Framework and a Methodology for Data Quality Assessment and Monitoring, Proceedings of the 12th International Conference on Information Quality, Cambridge, Massachusetts, 2007, pp. 333-346
[Batini et al. 2009]
Batini, C., Cappiello, C., Francalanci, C., Maurino, A., Methodologies for Data Quality Assessment and Improvement, in: ACM Computing Surveys, 41, 2009, No. 3, pp. 1-52
[Batini/Scannapieco 2006]
Batini, C., Scannapieco, M., Data Quality. Concepts, Methodologies and Techniques, Springer, Berlin, Deutschland et al., 2006
[Becker et al. 2003]
Becker, J., Holten, R., Knackstedt, R., Niehaves, B., Forschungsmethodische Positionierung in der Wirtschaftsinformatik – Epistemologische, ontologische und linguistische Leitfragen, Institut für Wirtschaftsinformatik Westfälische Wilhelms-Universität Münster, Münster, Deutschland, 2003
[Boisot/Canals 2004]
Boisot, M., Canals, A., Data, information and knowledge: have we got it right?, in: Journal of Evolutionary Economics, 14, 2004, No. 1, pp. 43-67
[Bourdreau/Couillard 1999]
Bourdreau, A., Couillard, G., Systems Integration and Knowledge Management, in: Information Systems Management, 16, 1999, No. 4, pp. 24-32
[Brinkkemper 1996]
Brinkkemper, S., Method Engineering: Engineering of Information Systems Development Methods and Tools, in: Information and Software Technology, 38, 1996, No. 4, pp. 275-280
[Bucher 2009]

Bucher, T., Ausrichtung der Informationslogistik auf operative Prozesse – Entwicklung und Evaluation einer situativen Methode, Dissertation, Institut für Wirtschaftsinformatik, Universität St. Gallen, St. Gallen, 2009

[Bund 2009]
Bund, Leistungs- und Finanzierungsvereinbarung, Bundesrepublik Deutschland, 2009

[Burnett et al. 1999]
Burnett, K., Ng, K. B., Park, S., A Comparison of the Two Traditions of Metadata Development, in: Journal of the American Society for Information Science, 50, 1999, No. 13, pp. 1209-1217

[Caballero et al. 2008]
Caballero, I., Calero, C., Piattini, M., Verbo, E., MMPro: A Methodology based on ISO/IEC 15939 to Draw up Data Quality Measurement Processes, Proceedings of the 13th International Conference on Information Quality, Cambridge, Massachusetts, 2008,

[Caballero et al. 2007]
Caballero, I., Verbo, E., Calero, C., Piattini, M., A Data Quality Measurement Information Model based on ISO/IEC 15939, Proceedings of the 12th International Conference on Information Quality, Cambridge, Massachusetts, 2007, pp. 393-408

[DAMA 2009]
DAMA, The DAMA Guide to the Data Management Body of Knowledge, Technics Publications, Bradley Beach, New Jersey, 2009

[DIN/EN/ISO 2000]
DIN/EN/ISO, Qualitätsmanagementsysteme: Leitfaden zur Leistungsverbesserung (ISO 9004:2000), DIN EN ISO 9004:2000 D, Deutsches Institut für Normung, 2000

[Ebner et al. 2011]
Ebner, V., Hüner, K. M., Otto, B., Fallstudie Bayer CropScience AG – Entwurf und Implementierung geschäftsorientierter Datenqualitätskennzahlen, BE HSG / CC CDQ / 23, Institut für Wirtschaftsinformatik Universität St. Gallen, St. Gallen, Schweiz, 2011

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

[Eppler/Helfert 2004]
Eppler, M. J., Helfert, M., A Classification and Analysis of Data Quality Costs, Proceedings of the 9th International Conference on Information Quality, Cambridge, Massachusetts, 2004, pp. 311-325

[Fan et al. 2008]

Fan, W., Geerts, F., Jia, X., Kementsietsidis, A., Conditional Functional Dependencies for Capturing Data Inconsistencies, in: ACM Transactions on Database Systems, 33, 2008, No. 2, pp. 1-48

[Frank 1997]
Frank, U., Erfahrung, Erkenntnis und Wirklichkeitsgestaltung. Anmerkungen zur Rolle der Empirie in der Wirtschaftsinformatik, in: Grün, O., Heinrich, L.J. (eds.), Wirtschaftsinformatik – Ergebnisse empirischer Forschung, Springer, Wien, Östereich, 1997, pp. 21-35

[Frank 2007]
Frank, U., Evaluation of Reference Models, in: Fettke, P., Loos, P. (eds.), Reference Modeling for Business Systems Analysis, Idea Group, Hershey, Pennsylvania et al., 2007, pp. 118-139

[Gregor 2006]
Gregor, S., The Nature of Theory in Information Systems, in: Management Information Systems Quarterly, 30, 2006, No. 3, pp. 611-642

[GS1 2010]
GS1, GS1 General Specifications, Version 10, Issue 1, Global Standards One, 2010

[Gutzwiller 1994]
Gutzwiller, T. A., Das CC RIM-Referenzmodell für den Entwurf von betrieblichen, transaktionsorientierten Informationssystemen, Physica, Heidelberg, Deutschland, 1994

[Heinrich et al. 2008]
Heinrich, B., Kaiser, M., Klier, M., Does the EU Insurance Mediation Directive Help to Improve Data Quality? A Metric-Based Analysis, Proceedings of 16th European Conference on Information Systems, Galway, Irland, 2008,

[Heinrich et al. 2007]
Heinrich, B., Klier, M., Kaiser, M., DQ metrics: A Novel Approach to Quantify Timeliness and its Application In CRM, Proceedings of the 12th International Conference on Information Quality, Cambridge, Massachusetts, 2007,

[Heinrich et al. 2009]
Heinrich, B., Klier, M., Kaiser, M., A Procedure to Develop Metrics for Currency and its Application in CRM, in: Journal of Data and Information Quality, 1, 2009, No. 1, pp. 1-28

[Hevner et al. 2004]
Hevner, A. R., March, S. T., Park, J., Ram, S., Design Science in Information Systems Research, in: Management Information Systems Quarterly, 28, 2004, No. 1, pp. 75-105

[Heym/Österle 1993]
Heym, M., Österle, H., Computer-Aided Methodology Engineering, in: Information and Software Technology, 35, 1993, No. 6/7, pp. 345-354

[Höning 2009]
Höning, F., Methodenkern des Business Engineering. Metamodell, Vorgehensmodell, Techniken, Ergebnisdokumente und Rollen, Dissertation, Institut für Wirtschaftsinformatik Universität St. Gallen, St. Gallen, Schweiz, 2009

[Huang et al. 1999]
Huang, K.-T., Lee, Y. W., Wang, R. Y., Quality Information and Knowledge, Prentice Hall, Upper Saddle River, New Jersey, 1999

[IEEE 1990]
IEEE, IEEE Standard Glossary of Software Engineering Terminology, lEEE Std 610.121990, The Institute of Electrical and Electronics Engineers, 1990

[IEEE 1998]
IEEE, IEEE Standard for a Software Quality Metrics Methodology, IEEE Std 1061-1998, The Institute of Electrical and Electronics Engineers, 1998

[ISO/IEC 2007]
ISO/IEC, ISO/IEC 15939. Systems and software engineering – Measurement process, ISO/IEC 15939:2007(E), International Organization for Standardization, 2007

[Kaiser et al. 2007]
Kaiser, M., Klier, M., Heinrich, B., How to Measure Data Quality? - A Metric-Based Approach, Proceedings of the 28th International Conference on Information Systems, Montreal, Kanada, 2007,

[Kaplan/Norton 1992]
Kaplan, R. S., Norton, D. P., The Balanced Scorecard: Measures that Drive Performance, in: Harvard Business Review, 70, 1992, No. 1, pp. 71-79

[Lee et al. 2006]
Lee, Y. W., Pipino, L. L., Funk, J. D., Wang, R. Y., Journey to Data Quality, MIT Press, Cambridge, Massachusetts, 2006

[Lee et al. 2002]
Lee, Y. W., Strong, D. M., Kahn, B. K., Wang, R. Y., AIMQ: a methodology for information quality assessment, in: Information & Management, 40, 2002, No. 2, pp. 133-146

[Loshin 2001]
Loshin, D., Enterprise Knowledge Management: The Data Quality Approach, Morgan Kaufmann, San Diego, California, 2001

[Loshin 2008]
Loshin, D., Master Data Management, Morgan Kaufmann, Burlington, Massachusetts, 2008

[March/Smith 1995]
March, S. T., Smith, G. F., Design and natural science research on information technology, in: Decision Support Systems, 15, 1995, No. 4, pp. 251-266

[McKinney Jr./Yoos 2010]
McKinney Jr., E. H., Yoos, C. J., Information About Information: A Taxonomy of Views, in: Management Information Systems Quarterly, 34, 2010, No. 2, pp. 329-344

[Mertens 2007]
Mertens, P., Integrierte Informationsverarbeitung 1: Operative Systeme in der Industrie, 16. Ed., Gabler, 2007

[Morgan/Krueger 1993]
Morgan, D. L., Krueger, R. A., When to use Focus Groups and why?, in: Morgan, D.L. (ed.), Successful Focus Groups, Sage, Newbury Park, California, 1993, pp. 3-19

[Nakatani et al. 2006]
Nakatani, K., Chuang, T.-T., Zhou, D., Data Synchronization Technology: Standards, Business Values and Implications, in: Communications of AIS, 17, 2006, No. 1, pp. 962-994

[Nuseibeh et al. 1996]
Nuseibeh, B. A., Finkelstein, Anthony, Kramer, J., Method Engineering for Multi-Perspective Software Development, in: Information and Software Technology, 38, 1996, No. 4, pp. 267-274

[Olle et al. 1988]
Olle, W. T., Hagelstein, J., Macdonald, I. G., Rolland, C., Sol, H. G., van Assche, F. J. M., Verrijn-Stuart, A. A., Information Systems Methodologies - A Framework for Understanding, Addison-Wesley, Wokingham, England, 1988

[Olson 2003]
Olson, J. E., Data Quality: The Accuracy Dimension, Morgan Kaufmann, San Francisco, California, 2003

[Österle 1995]
Österle, H., Business Engineering: Prozeß- und Systementwicklung, Band 1: Entwurfstechniken, 2. Ed., Springer, Berlin, Deutschland, 1995

[Österle/Blessing 2003]
Österle, H., Blessing, D., Business Engineering Modell, in: Österle, H., Winter, R. (eds.), Business Engineering, 2. Ed., Springer, Berlin, Deutschland, 2003, pp. 65-85

[Österle/Blessing 2005]
Österle, H., Blessing, D., Ansätze des Business Engineering, in: HMD – Praxis der Wirtschaftsinformatik, 42, 2005, No. 241, pp. 7-17

[Österle/Otto 2010]
Österle, H., Otto, B., Konsortialforschung: Eine Methode für die Zusammenarbeit von Forschung und Praxis in der gestaltungsorientierten Wirtschaftsinformatikforschung, in: WIRTSCHAFTSINFORMATIK, 52, 2010, No. 5,

[Österle et al. 2007]
Österle, H., Winter, R., Höning, F., Kurpjuweit, S., Osl, P., Business Engineering: Core-Business-Metamodell, in: Wisu – Das Wirtschaftsstudium, 36, 2007, No. 2, pp. 191-194
[Otto et al. 2010]
Otto, B., Hüner, K. M., Österle, H., A Cybernetic View on Data Quality Management, Proceedings of the 16th Americas Conference on Information Systems, Lima, Peru, 2010,
[Otto/Österle 2010]
Otto, B., Österle, H., Relevance Through Consortium Research? A Case Study, Proceedings of the European Conference on Information Systems, Pretoria, Südafrika, 2010,
[Peffers et al. 2008]
Peffers, K., Tuunanen, T., Rothenberger, M. A., Chatterjee, S., A Design Science Research Methodology for Information Systems Research, in: Journal of Management Information Systems, 24, 2008, No. 3, pp. 45–77
[Pipino et al. 2002]
Pipino, L. L., Lee, Y. W., Wang, R. Y., Data Quality Assessment, in: Communications of the ACM, 45, 2002, No. 4, pp. 211-218
[Pipino et al. 2005]
Pipino, L. L., Wang, R. Y., Kopcso, D., Rybold, W., Developing Measurement Scales for Data-Quality Dimensions, in: Wanf, R.Y., Pierce, E.M., Madnick, S.E., Fisher, C.W. (eds.), Information Quality. Advances in Management Information Systems, M.E. Sharp, Armonk, New York, 2005, pp. 37-51
[Price et al. 2008]
Price, R., Neiger, D., Shanks, G., Developing a Measurement Instrument for Subjective Aspects of Information Quality, in: Communications of AIS, 22, 2008, No. 1, pp. 49-74
[Price/Shanks 2005]
Price, R., Shanks, G., A semiotic information quality framework: development and comparative analysis, in: Journal of Information Technology, 20, 2005, No. 2, pp. 88-102
[Redman 1996]
Redman, T. C., Data Quality for the Information Age, Artech House, Boston, Massachusetts et al., 1996
[Redman 2005]
Redman, T. C., Measuring Data Accuracy. A Framework and Review, in: Wanf, R.Y., Pierce, E.M., Madnick, S.E., Fisher, C.W. (eds.), Information Quality. Advances in Management Information Systems, M.E. Sharp, Armonk, New York, 2005, pp. 21-36
[Rosemann/Vessey 2008]

Rosemann, M., Vessey, I., Toward Improving the Relevance of Information Systems Research to Practice: The Role of Applicability Checks, in: Management Information Systems Quarterly, 32, 2008, No. 1, pp. 1-22

[Scheer 1998]
Scheer, A.-W., ARIS – Vom Geschäftsprozess zum Anwendungssystem, 3. Ed., Springer, Berlin et al., 1998

[Schemm 2008]
Schemm, J. W., Stammdatenmanagement zwischen Handel und Konsumgüterindustrie: Referenzarchitektur für die überbetriebliche Datensynchronisation, Dissertation, Institut für Wirtschaftsinformatik, Universität St. Gallen, St. Gallen, Schweiz, 2008

[Schmidt 2010]
Schmidt, A., Entwicklung einer Methode zur Stammdatenintegration, Dissertation, Institut für Wirtschaftsinformatik Universität St. Gallen, St. Gallen, Schweiz, 2010

[Schütte/Rotthowe 1998]
Schütte, R., Rotthowe, T., The Guidelines of Modeling – An Approach to Enhance the Quality in Information Models, Proceedings of the 17th International Conference on Conceptual Modeling, Singapur, 1998, pp. 240-254

[Shankaranarayanan/Cai 2006]
Shankaranarayanan, G., Cai, Y., Supporting data quality management in decision-making, in: Decision Support Systems, 42, 2006, No. 1, pp. 302-317

[Shankaranarayanan et al. 2000]
Shankaranarayanan, G., Wang, R. Y., Ziad, M., IP-MAP: Representing the Manufacture of an Information Product, Cambridge, Massachusetts, 2000, pp. 1-16

[Shankaranarayanan et al. 2003]
Shankaranarayanan, G., Ziad, M., Wang, R. Y., Managing Data Quality in Dynamic Decision Environments: An Information Product Approach, in: Journal of Database Management, 14, 2003, No. 4, pp. 14-32

[Simon 1996]
Simon, H. A., The sciences of the artificial, 3. Ed., MIT Press, Cambridge, Massachusetts et al., 1996

[Smith/McKeen 2008]
Smith, H. A., McKeen, J. D., Developments in Practice XXX: Master Data Management: Salvation or Snake Oil?, in: Communications of the AIS, 23, 2008, No. 4, pp. 63-72

[Spiegler 2000]
Spiegler, I., Knowledge management: a new idea or a recycled concept?, in: Communications of the AIS, 3, 2000, No. 4, pp. 1-24

[Strong et al. 1997]
Strong, D. M., Lee, Y. W., Wang, R. Y., Data Quality in Context, in: Communications of the ACM, 40, 1997, No. 5, pp. 103-110
[vom Brocke 2003]
vom Brocke, J., Referenzmodellierung. Gestaltung und Verteilung von Konstruktionsprozessen, Logos Verlag, Berlin, Deutschland, 2003
[Wand/Wang 1996]
Wand, Y., Wang, R. Y., Anchoring Data Quality Dimensions in Ontological Foundations, in: Communications of the ACM, 39, 1996, No. 11, pp. 86-95
[Wang 1998]
Wang, R. Y., A product perspective on total data quality management, in: Communications of the ACM, 41, 1998, No. 2, pp. 58-65
[Wang et al. 1998]
Wang, R. Y., Lee, Y. W., Pipino, L. L., Strong, D. M., Manage Your Information as a Product, in: Sloan Management Review, 39, 1998, No. 4, pp. 95-105
[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, No. 4, pp. 5-33
[Weber 2009]
Weber, K., Data Governance-Referenzmodell: Organisatorische Gestaltung des unternehmensweiten Datenqualitätsmanagements, Dissertation, Institut für Wirtschaftsinformatik Universität St. Gallen, St. Gallen, Schweiz, 2009
[Weber et al. 2009]
Weber, K., Otto, B., Österle, H., One Size Does Not Fit All—A Contingency Approach to Data Governance, in: ACM Journal of Data and Information Quality, 1, 2009, No. 1, pp. 4:1-4:27
[Wedekind 2001]
Wedekind, H., Stammdaten, in: Mertens, P., Back, A., Becker, J., König, W., Krallmann, H., Rieger, B., Scheer, A.-W., Seibt, D., Stahlknecht, P., Strunz, H., Thome, R., Wedekind, H. (eds.), Lexikon der Wirtschaftsinformatik, 4. Ed., Springer, Berlin, Deutschland et al., 2001, p. 72

Appendix A Entity Types of the Method's Meta-model

Table A-1 describes the entity types of the method's meta-model (cf. section 5.1) and its relationships to entity types of the of the BE core meta-model (cf. section 2.1) as specified by HÖNING [2009].

Meta-Entity Type	Description
Data Object Attribute (short form: Attribute)	A data object attribute (short form: attribute) is the type of a data element. Data object attributes correspond to attributes in relational data models. Atomic values are assigned to atomic data object attributes when instantiated [Schmidt 2010]. **Relationships** Data elements can be instantiated from a data object attribute. A data object attribute is an integral part of a data object type.
Data Object	A data object is a set of several data elements which together form the instance of the respective data object type [Schmidt 2010]. **Relationships** A data object consists of several data elements. Data objects can be hierarchically structured. A data object represents a business object on the system layer. A data object is an instance of a data object type.
Data Object Type	A data object type describes a data object, e.g. on the basis of its range of values or its classification. It unites a set of related data object attributes which can be used to describe data objects (level of instances) abstractly (e.g. in a model) [Schmidt 2010]. **Relationships** Data objects can be instantiated from a data object type. A data object type consists of several data object attributes. A data object type realizes exactly one business object type by selecting specific object characteristics and mapping them onto data object attributes. Data object types can be hierarchically structured.
Data Quality	Data quality is a phenomenon perceived by a user of data elements. Data quality assesses the extent to which data elements meet the requirements of a data user [Wang/Strong 1996]. A data user can have different requirements of a (same) data element when performing different activities (i.e. context dependency of data quality). **Relationships** Data quality is a constituent part of process performance and is produced by support processes (e.g. data management, data quality management processes).
Data Quality Dimension	DQ dimensions represent specific aspects (e.g. believability, completeness, timeliness) of the phenomenon *data quality* [Wang/Strong 1996]. **Relationships** A DQ dimension is a special success factor. DQ dimensions conceptualize the phenomenon *data quality*. DQ dimensions can be hierarchically structured.
Data Quality	A DQ metric is a quantitative measure of data quality. A metrics system measures a metric

Meta-Entity Type	Description
Metric	by applying a measuring technique (e.g. verifying data elements with validation rules) at a measuring point (e.g. a database) at a point in time defined by a measuring frequency, the mapping of the measurement onto a measuring scale and the assignment of a unit of measure [IEEE 1998, DIN/EN/ISO 2000, Batini/Scannapieco 2006]. **Relationships** DQ metrics operationalize DQ dimensions. DQ metrics can be hierarchically structured.
DQM Initiative	A preventive DQM initiative (e.g. specification and use of DQ metrics) reduces the risk of data defects (reduction in the probability and/or the impact of a data defect). A reactive DQM initiative (e.g. duplicate elimination) removes data defects after they have occurred [Otto et al. 2010]. **Relationships** DQM initiatives can be hierarchically structured.
Business Object Attribute	A business object attribute (e.g. age, weight) denotes common characteristics of business object types. Business object attributes are characterized by attribute types and multiplicities [Schmidt 2010]. **Relationships** A business object attribute is an integral part of a business object type.
Business Object Type	Business object types (e.g. customers, products) unite a set of related business object attributes which can be used to describe real-world business objects (level of instances) abstractly (e.g. in a model) [Schmidt 2010]. **Relationships** A business object type is a collection of metadata forming a business object (instance). Several business objects can be instantiated from one business object type. A business object type consists of several business object attributes. A business object type is realized by exactly one data object type.
Metadata	Metadata is defined as data elements that describe the meaning as well as the business and technical characteristics of other data [Schmidt 2010]. Technical metadata (e.g. data types, field lengths, value ranges) specifies data objects (and their attributes) of an IS. Business metadata (e.g. responsibility for maintenance, correct process of data creation) describes the use of business objects (or the data objects representing them) in activities of an enterprise's organization structure. **Relationships** Metadata consists of special data elements. Metadata describes data object types. Metadata describes business object types.
Master Data Class	A master data class groups together master data objects of the same type (e.g. customer data, product data). **Relationships** A master data class is a special data object type. Master data classes can be hierarchically structured.
Master Data Object	A master data object is a special data object whose data elements change less frequently in comparison with other data objects (e.g. transaction data objects) [Schmidt 2010]. **Relationships**

Meta-Entity Type	**Description**
	A master data object is a special data object.

Table A-1: Definition of Meta-Entity Types of the Method

Appendix B Identification of Causal Chains

B.1 Examples of Causal chains

Table B-1 shows causal chains identified in case studies and focus group interviews. In each case, the context of a causal chain not only states the strategic objective, but also briefly describes the enterprise and the scope.

Causal Chain	Description
Fastening Technology / Customer Data	
Context	The company primarily sells its products directly to its customers (e.g. building contractors, tradesmen). This gives the company a good insight into current customer needs and consequently enables a customer-oriented innovation process. Equally, however, it calls for a flexible sales organization with a good global network as customer support can cover fewer customers than in the case of alternative distribution channels. The company's strategic objectives are therefore high customer satisfaction and a high level of trust in the services and products offered (achieved through consistency as far as possible in the allocation of customers and sales staff) and an efficient sales organization (i.e. good support for as many customers as possible by one sales staff member).
Business Problem	Alongside general business problems such as errors when compiling invoices, invoices sent to the wrong address or incorrectly planned central marketing initiatives, defects in customer data lead to special problems in direct selling. Defects in customer classification, for example, can mean that a customer is not actually offered products which would be of interest as the (perhaps new) member of the sales team is not fully aware of the customer's requirements. Moreover, defects in address and communication data have a direct impact on the (important – see above) effectiveness of a salesperson as he or she cannot optimally plan the selling route and first has to research the correct data.
Data Defect	Missing, inaccurate or outdated customer data and customer classification (i.e. assignment of customers to product groups which are potentially of interest).
Chemical Industry / Material Data	
Context	The corporation has consolidated systems from companies in over 120 countries into three regional ERP systems. An ongoing project is aimed at consolidating the three regional systems into one globally utilized ERP system. A central master data server distributes customer, supplier and material data to the regional systems and in future will distribute this data to the central ERP system. One of the corporation's strategic objectives is to consolidate material requirements on a group-wide basis so that better purchasing conditions can be negotiated by a central procurement department.
Business Problem	In production planning, defects in the product hierarchy are preventing the correct consolidation of requirements and consequently optimal requirements planning. When it comes to reporting, it is not possible to correctly assign the sales figures to the products and to the business units responsible.
Data Defect	Incorrect or incomplete classification of products with the product hierarchy (grouping of products in a tree structure), inconsistent classification of packed and unpacked materials.

Causal Chain	Description
Power Supply / Customer Data	
Context	Power Inc., see Table 5-1. One of the corporation's strategic objectives is to achieve high customer satisfaction so that customers are not lost to competitors (increased competition due to liberalization of the energy market).
Business Problem	Missing titles lead to incomplete letters, e.g. invoices or marketing information. Furthermore, complete titles are necessary so that personalized forms of address can be used (e.g. for welcome letters, requests from marketing).
Data Defect	Missing titles.
Business Problem	Billing blocks halt the dispatch of invoices among others if, for example, a recipient has gone abroad for a long period. However, billing blocks should not be set for too long without good reason as they slow up the billing process and cause additional costs.
Data Defect	Outdated billing blocks.
Business Problem	Invoices and reminders are sent to the wrong people. If customers want different addressees, this is explicitly noted in a check table.
Data Defect	Inconsistent address data for recipients of reminders and invoices.
Business Problem	An incorrect data element leads to the use of the wrong invoice forms and consequently to incorrect invoice printouts.
Data Defect	Incorrect data elements for invoice form selection (business and residential customers).
Business Problem	Customers receive a bill despite the fact that they have already stated *direct debit* (several times) as their chosen method of payment.
Data Defect	Outdated method of payment. If a customer reports that he/she is moving, i.e. his/her address is changing, the customer is retained as data object (business partner), but a new data object is created for billing data (contract account). In individual cases, the customer's chosen method of payment (e.g. direct debit, bill) is not transferred from the old contract account.
Consumer Goods Industry / Product Data	
Context	Consumer Goods Inc., see The corporation's strategic objectives are compliance with legal requirements, cost reduction and a high service level.
Business Problem	Defects in logistical data lead to additional effort and delays in logistics processes. For example, pallets weighing more than the maximum permitted weight for a warehouse cannot be placed in storage. This may require the pallets to be repacked, causing further delays and possibly incorrect deliveries.
Data Defect	Missing or inaccurate logistical product data (e.g. dimensions, weight, quantities).
Business Problem	Inconsistencies between the GTINs on products or logistical units (e.g. pallets) and the GTINs sent to a data pool lead to deliveries which cannot be handled smoothly by the customer. If inconsistencies are discovered prior to delivery of the goods, these lead to additional effort and possibly cost for the production of new artwork or new products.
Data Defect	Inconsistent GTINs.

Causal Chain	Description
Business Problem	It is not possible to release a bill of material (BOM) until all the required items on the list have been stated. The delay in releasing a bill of material leads to a late production start and a failure to meet delivery dates.
Data Defect	Incomplete BOM items.
Telecommunications Industry / Network Infrastructure Data	
Context	Telephone Inc., see Table 5-1. A strategic corporate objective is to win new customers. By using an availability check, potential new customers can find out if specific products are available at a specific address.
Business Problem	Defects in the network infrastructure data lead to incorrect answers to checks on whether specific products requiring the availability of broadband internet can be offered at a given address. False positive answers lead to customer dissatisfaction (ordered products are ultimately not available or not in an acceptable quality); false negative answers lead to lower sales revenues (products which are available are not actually offered).
Data Defect	Inconsistent network infrastructure data. Despite the fact that accuracy and completeness are actually the success factors for an error-free availability check, the complexity (scope, necessary level of detail) of the required documentation nonetheless makes the achievement of this goal unrealistic.
Watch Movement Production / Product Data	
Context	The company makes watch movements with a high proportion of manual production. In view of the high demand and limited production capacity, the company relies heavily on fast and, as far as possible, error-free production. The company's strategic objectives are therefore low variances in the start of production (key performance indicator for production planning), fast cycle times (time taken to produce a watch movement) and a low percentage of faulty movements (rejects).
Business Problem	If design drawings (exact specification of individual components of a watch movement) are missing at the start of production or if these are obviously outdated, they have to be requested from design engineers and retrospectively added to the bills of material by hand. This delays the start of production. Missing or outdated design drawings can also mean that drawings of similar products have to be used in production, which can lead to rejects.
Data Defect	Missing or outdated linkage of design drawings in bills of material.
Supplier to the Automotive Industry / Material Data	
Context	The company makes the same product parts at different plants on highly specialized production lines. Spare parts for these machines are usually expensive and have to be made to order, which means that production delays are possible. In view of the fact that there are similar production lines at various plants, it would in theory be possible to exchange stocked spare parts between plants. One of the company's strategic objectives is to enable spare parts for production lines to be exchanged between plants in order to cut the duration of maintenance work and reduce storage costs.
Business Problem	It is only possible to exchange suitable spare parts if these are correctly described and classified by technical data. Data defects lead to the identification of the wrong parts or to parts that are actually in stock not being found.
Data Defect	Missing or inaccurate technical data (e.g. dimensions, weight, classification, material) relating to spare parts for production lines.

Table B-1: Examples of Causal Chains

B.2 Guiding Questions for Identifying Critical Data Defects

The guiding questions given in Table B-2 support interviews for identifying business-critical data defects. They are structured according to the 15 DQ dimensions defined by WANG/STRONG [1996].

DQ Dimension	Possible Data Defect	Guiding Question
Access Security	Data elements for which change authorizations are supposed to have been given cannot be changed. Data elements for which no change authorizations are supposed to have been given can be changed.	Does data exist which you would like to change and are in fact supposed to be able to change, but cannot? Is it possible to directly add or correct missing or wrong data? Do you make changes to data which you are not in fact supposed to be able to do? Does data exist which you need but have no access to?
Accessibility	Obtaining the necessary data elements takes too long because system performance is inadequate or the route is too long.	Is the required data obtainable via the intended route (e.g. application used, stipulated display screens)? Can the required data be obtained from other sources (e.g. internet, telephone, application not envisaged for the process activity)? How often is this necessary?
Accuracy	Input forms contain a lot of mandatory fields where dummy values (i.e. *NULL* values) have to be entered due to a lack of knowledge about the effects. The available data elements are inadequate for distinguishing between two data objects without any doubt (duplicates). Different systems deliver different data elements for a specific set of facts (inconsistency).	Does it often occur that data is wrong (e.g. address data, wrong telephone numbers)? How do you know that the data is wrong? Are there input forms which contain many fields which you should complete but are unable to (e.g. due to lack of time or due to uncertainty)? Does it happen that data appears several times over although it should only exist once in the system (e.g. different addresses for the same customer)? Do data inconsistencies exist between different systems?
Appropriate Amount of Data	There are so many data elements that an employee has great difficulty finding the relevant data elements. It is necessary to obtain the necessary data elements via various routes (e.g. systems, screen forms, documents) in order to complete a single work step. It is not possible to enter all the necessary data elements, so e.g. user-defined text fields have to be (mis)used for this purpose.	Are there input forms or reports which contain such extensive data that you have trouble finding the relevant data? Are there work steps where you have to obtain or link data from several applications? Do you find that in specific work steps the input fields for data required for processing are missing? If so: Where exactly do you enter that data? Are there input forms where you have to enter a lot of information which does not seem to be relevant for the business transaction? How do you deal with this?
Believability	As data elements are wrong or inadequate in many cases, poor data quality is assumed from the outset.	Do you assume available data to be correct, or do you assume from the outset that it will contain errors?

DQ Dimension	Possible Data Defect	Guiding Question
	For that reason, data elements are requested which are in fact present.	Is specific data validated once again with other data (e.g. checking back with customers or research in alternative databases)? Which data does not seem believable to you? What makes you think that?
Completeness	The available data elements are inadequate for completing a specific work step. It is not possible to obtain the missing data through other sources (e.g. telephone, other software). Data elements are missing.	Are data elements which should be present in specific screen forms or forms but are actually missing? Are there data elements which are shown in a way that is not easy to understand or are too severely abbreviated?
Concise Representation	The representation of data elements is not helpful for processing a specific task and could be more precise.	Are there cases where you are offered data elements that you first have to think about for a long time before you understand them syntactically, or where you have to use additional sources of information?
Ease of Understanding	Data elements are readable (i.e. their meaning is intrinsically clear), but it is unclear how they are to be used or why they are offered.	Is there data which you understand syntactically but are unclear about its meaning and/or impact?
Interpretability	Data elements are not understandable due to an unknown language or incorrectly used coding.	Is there data which is shown in what for you is a foreign language? Is there data which shows symbols (e.g. “$%§?#”) where this should not be the case? Are there fields which show incomprehensible values (“gibberish”)?
Objectivity	Data elements deliver the opinion or appraisal of individuals and not objective information.	Do you doubt the correctness of data because you suspect a subjective opinion or appraisal, but actually expect objective information?
Relevancy	So many data elements are offered for a business transaction that selecting the data elements required for the business transaction involves considerable effort.	Are you “swamped” with data, i.e. do you find it difficult to select the data required to process a task?
Representational Consistency	No uniform data format, e.g. of address data in different applications. This can lead to misunderstandings or technical problems with data exchange. Different systems access different databases and provide different representations of data elements in the same context. In some cases, data elements have not been entered in the envisaged input fields and are therefore shown in the wrong place.	Is there data which is entered or shown differently in different applications? Are there cases where different data is shown in respect of one customer in different systems? Are there screen forms which show information “the wrong way round” (e.g. surname in the field for first name)?
Reputation	Available data (i.e. the application or department that delivers the data) is not trusted.	Do you doubt the correctness of data because of its origin? Do you regularly crosscheck specific data against

DQ Dimension	Possible Data Defect	Guiding Question
		comparable information from other applications or departments?
Timeliness	Data elements are outdated despite the fact that more current data elements are available or have actually been entered. Reworking or storing data elements takes so long that the impression arises that the input has not yet taken place. For that reason, data is entered several times (possible cause of duplicates).	Is input data available or updated on time for subsequent tasks? Are there cases where repeated entries are made as it is only visible after a delay, if at all, that these have actually been performed?
Value Added	Despite being correct, data elements are not helpful for processing the respective task.	Is there data which you understand but which does not help you when processing the respective task?

Table B-2: Guiding Questions for Identifying Critical Data Defects

Appendix C Requirements to be met by DQ metrics

The requirements to be met by DQ metrics described in Table C-3 were identified in a focus group interview (June 11, 2008, 90 minutes, 18 participants from 10 different companies).

Requirement	Description
Cost / Benefit	The ratio of the costs required to define and measure the values of a DQ metric to the benefit obtained (controlled error potential) should be positive (< 1).
Definition of Measuring Frequency	For each DQ metric, the times at which values are to be measured should be defined.
Definition of Measuring Point	For each DQ metric, the point at which values are to be measured (e.g. database, process, department) should be defined.
Definition of Measuring Technique	For each DQ metric, the tool (e.g. questionnaire, software) to be used to measure its value should be defined.
Definition of Scale	For each DQ metric, a scale (e.g. percent, grades, time) should be defined.
Limitation of the Data to be Validated	For each DQ metric, the data (e.g. material master, European customers) to be analyzed should be defined.
Escalation Process	For each DQ metric, activities should be defined in relation to a specific threshold, i.e. from what value which activity is to be carried out.
Area of Validity	For each DQ metric, an area of validity should be defined in which its values apply.
SMART Principles	A DQ metric should fulfill the SMART goals (specific, measurable, achievable, realistic, timely).
Disturbance Variables	The metadata of a DQ metric should contain information on possible disturbance variables, i.e. describe possible events or influences which might bias the values of the DQ metric.
Responsibility	For each DQ metric, clear responsibilities should be defined, i.e. which values of the DQ metric are to be reported to whom, who is responsible for maintaining/updating the metric (e.g. currency/purpose of definition, implementation of measuring technique).
Comparability	A DQ metric should be defined so that its values can be compared with other metrics (DQ metrics and process metrics).
Ease of Understanding	For each DQ metric, metadata should be available which explain its intended purpose and the correct interpretation of its values.
Use in SLAs	DQ metrics should be suitable for use in service level agreements.
Visualization	It should be possible to visualize the values of a DQ metric (e.g. time lines, graphs).
Repeatability	It should be possible to measure the values of a DQ metric not just once, but multiple times.
Target Value	For each DQ metric, a target value should be defined.
Assignment of DQ Dimension	It should be possible to assign a DQ metric to one or more DQ dimensions.
Assignment of	It should be possible to assign a DQ metric to a concrete (enterprise-specific) busi-

Requirement	Description
Business Problem	ness problem.
Assignment of Process Metrics	It should be possible to assign a DQ metric to one or more process metrics.
Assignment of Corporate Strategy	It should be possible to assign a DQ metric to one or more strategic objectives of the enterprise.

Table C-3: Requirements to be Met by DQ Metrics

Appendix D Examples of Validation Rules

Table D-4 shows an extract of the validation rules which were defined by applying the technique *specification of validation rules* (cf. section 5.2.5.1) in participative case studies with Consumer Goods Inc. and Power Inc. Only the validations performed for a data object d are described in each case. The selection of relevant (i.e. potentially defective) data objects (cf. sections 5.2.2 and 5.2.5) for the respective validation as well as an iterative loop (cf. Figure 5-8) for these selected data objects is assumed.

DQ Dimension / Rule	Description
Currency / Timeliness	

`if (CurrentTime() + threshold) > FirstDemand(d)`
`    if d.BoM.Status != 'released' : return 0`
`end if`
`return 1`
 	
`CurrentTime()` returns the current time.
`FirstDemand(d)` returns the point in time at which data element `d` is used for the first time (e.g. start of production).
	Verifies whether the release of a bill of material (BoM) has been performed in a timely manner (defined by the *threshold*).

`if (CurrentTime() + threshold) > FirstDemand(d)`
`    if d.CommodityCode   == NULL : return 0`
`    if d.CountryOfOrigin == NULL : return 0`
`end if`
`return 1`
	Verifies whether the data required for customs clearance is available on time.

`app1 = LastChange(A1.d.BoM.Status, 'released')`
`app2 = LastChange(A2.d.BoM.Status, 'released')`
`if app1 > app2 + threshold : return 0`
`return 1`
 	
`LastChange(a,e)` returns the point in time at which data element `e` was last assigned to attribute `a`.
	Verifies whether the release of a BOM in application A1 was also performed in a timely manner in application A2 (actually a consistency check, but classified here as the point in time is verified).

Change Frequency	

`if d.BoM.Status == 'released'`
`    for each position p in d.BoM`
`        if NumberOfChanges(p) > 1 : return 0`
`    end for`
`end if`
`return 1`
 	
`NumberOfChanges(a)` returns the number of changes to a data element of attribute `a` since initial input of the data element.
	Verifies how often the items of a BOM have been changed prior to its release. Frequent changes are an indication of problems with the data maintenance process and therefore an indication of data defects.

`if NumberOfChanges(d.BoM.Status) > 1 : return 0` `return 1`	Verifies how often the status of a BOM has been changed. The frequent resetting of an article that has already been released (representa-

DQ Dimension / Rule	Description
	tion of a product in the system) is an indication of problems with the data maintenance process and therefore an indication of data defects
Accuracy	

`for each data object d2`
`if d.GTIN == d2.GTIN`
`if GS1(d,d2) == FALSE : return 0`
`end if`
`end for`
`return 1`
 	
`GS1(d1,d2)` returns `TRUE` if the weights of two data objects `d1` and `d2` do not vary by more than 20%, otherwise `FALSE`.
	Verifies whether two data objects with the same GTIN have the characteristics required by GS1 *no weight variance of more than 20% for the same GTIN*.

`for each data object d2`
`if d.GTIN == d2.GTIN`
`if d.SH.Pieces != d2.SH.Pieces : return 0`
`if d.PU.Pieces != d2.PU.Pieces : return 0`
`if d.PA.Pieces != d2.PA.Pieces : return 0`
`end if`
`end for`
`return 1`
	Verifies whether two data objects with the same GTIN have the characteristics required by GS1 *same quantity on all logistical layers for the same GTIN* (SH: shrink; PU: packaging unit; PA: pallet).

`for each data object d2`
`if EditDist(d,d2) > threshold : return 0`
`return 1`
`EditDist(s1,s2)` returns the minimum edit distance between two strings `s1` and `s2`.
	Verifies whether two data objects are so similar that a duplicate is to be assumed. Validation by means of the minimum edit distance calls for character sequences as input and is therefore only suitable for certain data object types (e.g. address data).

Consistency	

`if d.GTIN != ArtworkGTIN(d) : return 0`
`return 1`
 	
`ArtworkGTIN(d)` returns the GTIN incorporated in a product artwork (if necessary, by applying a text recognition algorithm).
	Verifies the consistency of the GTIN depicted in a product artwork and the GTIN stored in the system.

`for each local system l`
`if d.GTIN != l.d.GTIN : return 0`
`if d.SH.GTIN != l.d.SH.GTIN : return 0`
`if d.VE.GTIN != l.d.PU.GTIN : return 0`
`if d.PA.GTIN != l.d.PA.GTIN : return 0`
`end for`
`return 1`
	Verifies the consistency of logistical data which is distributed to local systems by a central system, and transferred from there to a GS1-compliant data pool.

`for each business partner b`
`pm = NULL // payment method`
`for each account a of b.Accounts`
`if pm == NULL : pm = a.PaymentMethod`
`else if pm != a.PaymentMethod : return 0`
`end for`
`end for`
`return 1`
	Verifies whether the same method of payment is stated for all contract accounts of a customer. Exactly one contract account is always assigned to one customer. Several accounts for one customer only occur if the customer moves (change of address) and the old account is deactivated (but not deleted). If the customer's

DQ Dimension / Rule	Description
	preferred method of payment is not transferred, the deactivated contract account receives the more current method of payment.
Completeness	

`if d.GrossWeight == NULL : return 0`
`if d.Height      == NULL : return 0`
`if d.Length      == NULL : return 0`
`if d.Width       == NULL : return 0`
`if d.Volume      == NULL : return 0`
`return 1`
	Verifies the completeness of logistical data for a product.

`if d.FirstName1 == NULL : return 0`
`if d.Surname    == NULL : return 0`
`if d.Salutation == ‘Mr and Mr’ OR ‘Mr and Ms’ OR ‘Ms and Ms’`
`    if d.FirstName2 == NULL : return 0`
`    if d.Surname    == NULL : return 0`
`end if`
`return 1`
	Verifies whether first name and surname and – if required by the data element stored as title – also a second first name and surname are stated for a business partner.

Table D-4: Examples of Validation Rules

Appendix E Documentation Templates

E.1 Templates for Activity I.1

Method for the Specification of Business-oriented Data Quality Metrics

Activity I.1: Define the scope — Documentation template *Sponsor Interview*

Roles

Corporate Data Steward

Name, contact information

Sponsor

Name, contact information

Finished Projects

Projects having discussed data defects or DQM initiatives

Running Projects

Projects currently discussing data defects or DQM initiatives

Future Projects

Planned projects addressing data defects or DQM initiatives

Known Business Problems

Known business problems which potentially are caused by data defects

University of St. Gallen, Institute of Information Management
Chair of Prof. Dr. Hubert Österle

Figure E-1: Documentation Template for Sponsor Interview

Method for the Specification of Business-oriented Data Quality Metrics

Activity I.1: Define the scope — Documentation template *Scoping*

Roles

Corporate Data Steward	Name, contact information
Process Owner	Name, contact information

Business Process or Activity

Identifier	Name, ID
Description	Process outcome (e.g. products or services), strategic impact for the company

Prozesskennzahl

Identifier	Name, ID
Responsibility	Name, contact information
Description	Relation to business, scale, unit of measurement, measuring frequency, target value

Business Problem

Identifier	Name, ID
Description	Problem frequency, caused costs or other effects, impaired business goals
Data Defect as potential cause	Data defect which potentially cause the business problem

University of St. Gallen, Institute of Information Management
Chair of Prof. Dr. Hubert Österle

Figure E-2: Documentation Template for Scoping

E.2 Templates for Activity I.2

Method for the Specification of Business-oriented Data Quality Metrics

Activity I.2: Identify data and IT systems | Documentation template *Data Storage and Constraints*

Roles

Corporate Data Steward	Name, contact information
Process Owner	Name, contact information

Constraints

Identifier	Name, ID
Business Problem	Identifier, s. documentation template *Scoping*
Data Object Type	Name, e.g. customer data, material data
Data Storage	Identifier
Responsibility	Name, contact information
Constraints	• • • • • • • • • • • Only active data objects, only data for certain products, only customers of a certain age, …

University of St. Gallen, Institute of Information Management
Chair of Prof. Dr. Hubert Österle

Figure E-3: Documentation Template for Data Storage and Constraints

Methode zur Spezifikation geschäftsorientierte Datenqualitätskennzahlen

Aktivität I.2: Daten und IT-Systeme identifizieren Dokumentationsvorlage *Leitfragen*

Rollen

Konzern-Datensteward	Name, Kontaktdaten
Prozess-eigentümer	Name, Kontaktdaten

Leitfragen

DQ-Dimension	Aktualität, Änderungshäufigkeit, Genauigkeit, Konsistenz, Objektivität, Rechtzeitigkeit, Vollständigkeit, …
Leitfragen	• •
DQ-Dimension	Aktualität, Änderungshäufigkeit, Genauigkeit, Konsistenz, Objektivität, Rechtzeitigkeit, Vollständigkeit, …
Leitfragen	• •
DQ-Dimension	Aktualität, Änderungshäufigkeit, Genauigkeit, Konsistenz, Objektivität, Rechtzeitigkeit, Vollständigkeit, …
Leitfragen	• •
DQ-Dimension	Aktualität, Änderungshäufigkeit, Genauigkeit, Konsistenz, Objektivität, Rechtzeitigkeit, Vollständigkeit, …
Leitfragen	• •

Universität St. Gallen, Institut für Wirtschaftsinformatik
Lehrstuhl Prof. Dr. Hubert Österle

Figure E-4: Documentation Template for Guiding Questions for DQ Dimensions

E.3 Templates for Activity I.3

Method for the Specification of Business-oriented Data Quality Metrics

Activity I.3: Identify causal chains — Documentation template *Causal Chains*

Roles

Corporate Data Steward	Name, contact information
Business Data Steward(s)	Name, contact information
Data User	Name, contact information
Technical Data Steward(s)	Name, contact information
Process Owner	Name, contact information

Business Problem

Identifier	Name, ID
Process Activity	Identifier
Impact	Which business goals are impaired? Frequency, impact (e.g. workload, defective products)

Data Defect

Identifier	Name, ID
Data Objects	Identifier (s. *Constraints*, activity I.2)
Data Quality Dimension	s. documentation template *Guiding Questions*
Causation	How does the data defect cause the business problem? Always? Circumstances? Constraints?

University of St. Gallen, Institute of Information Management
Chair of Prof. Dr. Hubert Österle

Figure E-5: Documentation Template for Causal Chains

E.4 Templates for Activity II.1

Method for the Specification of Business-oriented Data Quality Metrics

Activity II.1: Define and prioritize requirements — Documentation template *Requirements*

Roles

Corporate Data Steward	Name, contact information
Business Data Steward(s)	Name, contact information
Sponsor	Name, contact information
Technical Data Steward(s)	Name, contact information

Requirements

Identifier	Name, ID	Priority	
Description			
Identifier	Name, ID	Priority	
Description			
Identifier	Name, ID	Priority	
Description			
Identifier	Name, ID	Priority	
Description			

University of St. Gallen, Institute of Information Management
Chair of Prof. Dr. Hubert Österle

Figure E-6: Documentation Template for Requirements

E.5 Templates for Activity II.2

Method for the Specification of Business-oriented Data Quality Metrics

Activity II.2: Specify data quality metrics — Documentation template *Validation Rules*

Roles

Corporate Data Steward	Name, contact information
Business Data Steward(s)	Name, contact information
Technical Data Steward(s)	Name, contact information

Validation Rule

Identifier	Name, ID
Business Problem	Identifier, see documentation template *Causal Chains*
Data Defect	Identifier, see documentation template *Causal Chains*
Attributes	• • •
Possible Validation Procedure	• • • • • • • Which characteristics of the data attributes (e.g. only certain data elements, formats) constitute the defect?

Figure E-7: Documentation Template for Validation Rules

Method for the Specification of Business-oriented Data Quality Metrics

Activity II.2: Specify data quality metrics — Documentation template *Data Quality Metric*

Roles

Corporate Data Steward	Name, contact information
Business Data Steward(s)	Name, contact information
Technical Data Steward(s)	Name, contact information

Data Quality Metric

Identifier	Name, ID
Data Quality Dimension	See documentation template *Guiding Questions, Causal Chains*
Geschäftsprobleme	• • • Identifier, see documentation template *Causal Chains*
Measuring Point	E.g. a data storage, see documentation template *Data Storage and Constraints*
Calculation Formula	Mapping of measurements by validation rules on a scale, assignment of a unit of measure
Validation Rules	• • • • • Identifier, see documentation template *Validation Rules*

University of St. Gallen, Institute of Information Management
Chair of Prof. Dr. Hubert Österle

Figure E-8: Documentation Template for DQ Metrics

E.6 Templates for Activity III.1

Method for the Specification of Business-oriented Data Quality Metrics

Activity III.1: Check the fulfillment of requirements — Documentation template *Requirements Validation*

Roles

Corporate Data Steward	Name, contact information
Business Data Steward(s)	Name, contact information
Technical Data Steward(s)	Name, contact information

Requirements Validation

Identifier	See documentation template *Requirements*	Approved?	
Arguments	E.g. reason why a requirement is not fulfilled, but the specification is approved however.		
Identifier	See documentation template *Requirements*	Approved?	
Arguments	E.g. reason why a requirement is not fulfilled, but the specification is approved however.		
Identifier	See documentation template *Requirements*	Approved?	
Arguments	E.g. reason why a requirement is not fulfilled, but the specification is approved however.		
Identifier	See documentation template *Requirements*	Approved?	
Arguments	E.g. reason why a requirement is not fulfilled, but the specification is approved however.		

University of St. Gallen, Institute of Information Management
Chair of Prof. Dr. Hubert Österle

Figure E-9: Documentation Template for Requirements Validation

E.7 Templates for Activity III.2

Method for the Specification of Business-oriented Data Quality Metrics

Activity III.2: Document metrics specification | Proposal for documentation structuring

Introduction

- Known data defects and business problems caused by these defects.
- What are potential benefits (e.g. more accurate demand planning, improved decission support) of increased data quality?

Fundamental Concepts

Data Quality Management

- Distinction of preventive and reactive DQM initiatives.
- Motivation of data quality metrics as a tool for preventive DQM.

Data Quality Measurement

- Necessity of aligning data quality metrics with the business goals of a company.
- Explanation of concepts *Causal Chain* and *Business-oriented Data Quality Metric*.

Results Causal Chain Identification

- List of identified business problems and data defects, structured as causal chains.
- Visualization of impact ratings of identified business problems (e.g. heat map).
- Description of business problems addressed by specified data quality metrics.

Data Quality Metrics

Metric 1

- Short description with identifier and addressed business problem.
- Addressed data defects.
- Measuring point and data constraints.
- Validation rules.

Metric 2

…

Structuring of Measuring System

- Description with rationales of selected calculation formula.
- Mapping of validation rules to data quality metrics.

Data Quality Simulation

- Quantification of data quality risks monitored by specified data quality metrics.
- Simulation of several scenarios and explanation of the impact of data defects on the data quality metrics values.

Outline for Implementation

- Prioritization of specified data quality metrics with respect to expected effort for implementation.
- Design draft for a measuring system (e.g. cockpit, scorecard, dashboard) for visualizing data quality measurements.

University of St. Gallen, Institute of Information Management
Chair of Prof. Dr. Hubert Österle

Figure E-10: Proposal for Documentation Structuring

Appendix F Focus Group and Guiding Questions for Interview

On February 9, 2010, 11 subject matter experts from 10 multinational companies (see Table F-1), each with 3 to 20 years‘ experience in master data management and DQM, evaluated the method in a focus group interview. After a 45-minute presentation of the method, the panel was asked to discuss and evaluate the method with the aid of several guiding questions (see Table F-2).

Company	Function
Bayer CropScience AG	Integration Manager Enterprise Master Data Management
Beiersdorf AG	Head of Supply Chain Data Process Management
Beiersdorf AG	Head of Team Beiersdorf Shared Services Master Data
DB Netz AG	Strategic Infrastructure Data Management, Railway Geographical Data
Festo AG & Co. KG	Head of Product Lifecycle Management
Grundfos A/S	Supply Chain Coordinator
Henkel AG & Co. KGaA	Project Manager Data Quality Strategy
Metro Group AG	Head of Business Intelligence Support
Rewe-Zentral AG	Project Manager Data Quality
Siemens Enterprise Communications GmbH & Co. KG	Vice President Master Data
Syngenta Crop Protection AG	Lead Steward Material Master Data

Table F-1: Subject Matter Experts of Focus Group Interview (February 9, 2010)

Relevance perspective
• Does the method‘s objective (i.e. identifying causalities between data defects and business problems, and specifying data quality metrics based on these causalities) address a relevant task from a company‘s point of view? • Does the method produce a relevant result? • Do the method‘s output (i.e. the metrics) serve the intended purpose?
Economic perspective
While the method aims at reducing costs (e.g. by providing a structured procedure, examples, and interview guidelines), the method‘s configuration and application cause costs, too. • Is the effort caused by the method‘s configuration and application acceptable, or would any ad-hoc approach serve the purpose as well? • Do the benefits of the method application exceed its costs?
Deployment perspective
In order to foster communication between the stakeholders involved, the method should be understandable. It should correspond to concepts, the prospective method users are familiar with.

• Are the concepts of the method (i.e. the meta-model, the procedure model) understandable? • Will the approach be accepted in a company? Is there a risk of a ―not invented here‖ syndrome?
Engineering perspective
• Does the method fulfill the requirements to be taken into account? • Is the method‘s specification and structuring suited for supporting the intended purposes? • Does the method describe the considered aspects correct and in a sufficient depth? • Is the method‘s structuring clear and understandable?
Epistemological perspective
• Is the method design process clear and understandable? Are there method elements (e.g. a role, an activity) without an obvious purpose? • Does the method design process generate credibility of the method, or does the method appear to be an ad hoc approach?

Table F-2: Guiding Questions for Focus Group Interview (February 9, 2010)