8-2010

A Cybernetic View on Data Quality Management

Boris Otto
University of St. Gallen, boris.otto@tu-dortmund.de

Kai M. Hüner
University of St. Gallen, kai.huener@unisg.ch

Hubert Österle
University of St. Gallen, hubert.oesterle@unisg.ch

A Cybernetic View on Data Quality Management

Boris Otto
University of St. Gallen,
Institute of Information Management
boris.otto@unisg.ch

Kai M. Hüner
University of St. Gallen,
Institute of Information Management
kai.huener@unisg.ch

Hubert Österle
University of St. Gallen,
Institute of Information Management
hubert.oesterle@unisg.ch

ABSTRACT

Corporate data of poor quality can have a negative impact on the performance of business processes and thereby the success of companies. In order to be able to work with data of good quality, data quality requirements must clearly be defined. In doing so, one has to take into account that both the provision of high-quality data and the damage caused by low-quality data brings about considerable costs. As each company's database is a dynamic system, the paper proposes a cybernetic view on data quality management (DQM). First, the principles of a closed-loop control system are transferred to the field of DQM. After that a meta-model is developed that accounts for the central relations between data quality, business process performance, and related costs. The meta-model then constitutes the basis of a simulation technique which aims at the explication of assumptions (e.g. on the effect of improving a data architecture) and the support of DQM decision processes.

Keywords

Cybernetics, data quality management, data quality costs, design science research, simulation.

INTRODUCTION

High-quality data is a prerequisite for companies to meet the requirements posed by numerous strategic business drivers, such as global process harmonization, integrated customer management, global spend analysis, or compliance with legal provisions and laws. Companies use data when performing their everyday business operations, i.e. to produce goods, to render services, or to do research and development. The quality of the data used affects business processes (e.g. with regard to lead time) as well as the outcome of business processes (e.g. customer satisfaction, product quality). In small-batch production, for example, fast availability of construction drawings at different steps of the manufacturing process as well as completeness and correctness of parts lists are crucial prerequisites for short process lead times. And in the services industry (in a call center, for example) fast availability of all data about each customer (e.g. previous orders and contracts, complaints) is mandatory for short processing times and high customer satisfaction. In order to be able to work with data of good quality, data quality requirements must clearly be defined. In doing so, one has to take into account that both the provision of high-quality data and the damage caused by low-quality data brings about considerable costs.

However, studies about the relation between the effort required to achieve a certain degree of data quality and the costs caused by this effort are missing. Lee et al. (Lee, Pipino, Funk and Wang, 2006) do propose the use of real options (Black and Scholes, 1973) for cost-benefit analysis in the field of data quality management (DQM) as well as simulation as a technique to determine the value of such options (Amram and Kulatilaka, 1998). However, these reflections have not been further specified so far.

As each company's database is a dynamic system, the paper proposes a cybernetic view on DQM. On the basis of a meta-model a simulation technique is developed, which does not intend to allow for automatic control of DQM initiatives or forecasting of the development of data quality over time, but which aims at the explication of assumptions (e.g. expected effect of a data clearing measure) and the support of decision processes in corporate DQM. Furthermore, the simulation technique supports the specification process of business oriented data quality metrics to identify data defects that are considered to have a mission critical impact (Otto, Hüner and Österle, 2009). When searching for causal relations (i.e. a specific data defect causes a specific business problem which jeopardizes a specific strategic company goal) underlying such metrics, scenarios may be created and simulated to verify the plausibility of assumptions made.

The paper starts with a short presentation of related work. Then the underlying research methodology is presented. After that the paper introduces a meta-model as a basis for a DQM simulation technique. To facilitate scenario modeling the paper specifies various impact patterns representing relations between various elements of the model plus a simulation model. Both the meta-model and the simulation technique are then evaluated in a business scenario at an international telecommunications company. Finally the paper concludes with a summary and an outlook on future research to be done.

STATE OF THE ART

Measuring Data Quality

Data quality has been a research topic in numerous empirical (Wang and Strong, 1996), pragmatic (English, 1999; Redman, 1996) and theoretical (Price and Shanks, 2005; Wand and Wang, 1996) studies, the outcome of which mainly were lists and categories of data quality dimensions. Many of these studies have in common that data quality is seen as something that is mainly determined by the data's '*fitness for use*', i.e. whether data is of good or poor quality depends on the context it is used in and the user it is used by. While the quality of a certain set of customer data, for example, can be sufficient to conduct a product marketing initiative (having available correct and complete e-mail addresses), it may not be good enough for duplicate free consolidation with another set of customer data (as an unambiguous key is missing and address data are inconsistent). Apart from the structural conceptualization of data quality by means of various data quality dimensions, the measurement of data quality (i.e. of data quality dimensions) has been a central issue of many scientific studies (Burgess, Gray and Fiddian, 2004; Gebauer, Caspers and Weigel, 2005; Gustavsson, 2006; Heinrich, Klier and Kaiser, 2009; Naumann and Rolker, 2000). In this respect, the impact of data defects on the operations of companies needs to be taken into account when data quality metrics are being designed (Batini, Barone, Mastrella, Maurino and Ruffini, 2007; Caballero, Calero, Piattini and Verbo, 2008; Otto et al., 2009).

Costs and Benefits of Data Quality (Management)

A number of studies has dealt with the impact data defects have on business operations (Eppler and Helfert, 2004; Fisher and Kingma, 2001; Joshi and Rai, 2000), introducing various classifications of costs caused by DQM. In this respect, it is important to note that both the provision of high-quality data and the damage caused by low-quality data brings about considerable costs. For example, a clockwork manufacturer's production process involves both automated and manual sub-processes, the latter comprising activities such as assembly of special pointers for certain product series and brands. If for such a manual process an assembly instruction drawing is not available for some reason (from an information systems perspective: the linking of an electronic document with the parts list of a production order), process lead time gets longer while process performance per time unit (e.g. the number of clockworks produced per hour) gets lower. Also, outdated or false construction drawings can lead to production of scrap parts, leading also to decreased process performance. Potential measures to ensure good availability of data relevant in the production process may include regular verification of assembly and construction drawings or training of constructing engineers to raise awareness about the effects of missing or false drawings.

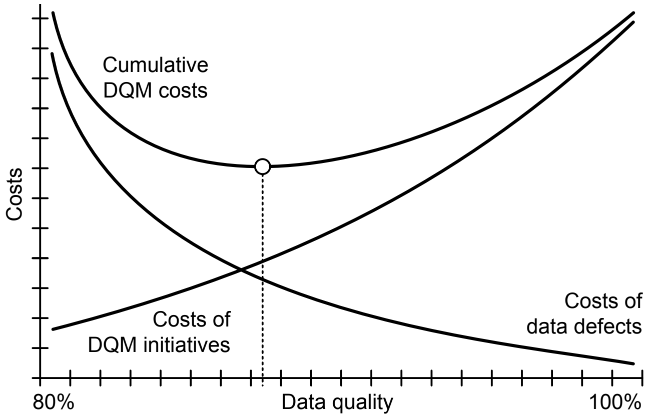

Figure 1. Cost-Benefit-Perspective on DQM

Obviously, data defects (here: a missing link) can cause business problems (here: missing construction drawings), which in turn lead to increased costs (here: production of less clockworks per time unit). DQM initiatives could help prevent or counteract data defects (here: through manual linking of construction drawings prior to the production process or through implementation of validation systems in preceding processes). So while DQM initiatives can reduce overall costs, they also cause costs for their development and implementation. Therefore, the objective is to determine the minimum of cumulated costs taking into account costs caused by data defects on the one hand and costs of DQM initiatives on the other hand (Eppler et al., 2004). Figure 1 shows a cost-benefit perspective on DQM.

RESEARCH METHODOLOGY

The meta-model together with the DQM simulation technique is an outcome of design oriented research, following Design Science Research as a methodological paradigm. Design Science Research aims at the design of artifacts (i.e. constructs, models, methods and instantiations) supposed to solve practical problems (Hevner, March, Park and Ram, 2004). In this context, constructs – similar to a language – define concepts for describing a specific problem, while models use constructs for describing concrete problems and possible solutions. The meta-model presented in this paper is a Design Science Research construct defining concepts for the description of simulation models which in turn describe concrete DQM scenarios by formulating assumptions regarding the impact of data defects and DQM initiatives as probabilistic variables.

Taking into account various requirements regarding the design of artifacts (Hevner et al., 2004), the design process for each artifact consists of several phases, namely problem identification, definition of design objectives, artifact design, artifact evaluation, and – if necessary – artifact redesign (Peffers, Tuunanen, Rothenberger and Chatterjee, 2008). The research context of the design process of the method is constituted by the Competence Center Corporate Data Quality. The Competence Center is part of the Business Engineering research program at the University of St. Gallen. Since 2006 it has been developing – together with renowned partner companies – solutions for support of quality management of corporate data.

SIMULATION APPROACH

Data Quality Management as a Closed-Loop Control System

Each company's database can be considered as an open system characterized by the quality of the data changing over time (Orr, 1998). Data quality is affected by both internal factors (e.g. no standard procedures in the use of data, no validation procedures concerning the identification of duplicates in the data collection process) and external factors (e.g. customer address data affected by customers who have moved to another place). As data quality shows a dynamic behavior over time and also because there has to be a balancing between the costs for DQM on the one hand and the costs expected to occur from poor data quality on the other hand, a systems theory approach can be applied. Figure 2 shows the interrelations described from a systems theory perspective (Lunze, 2008).

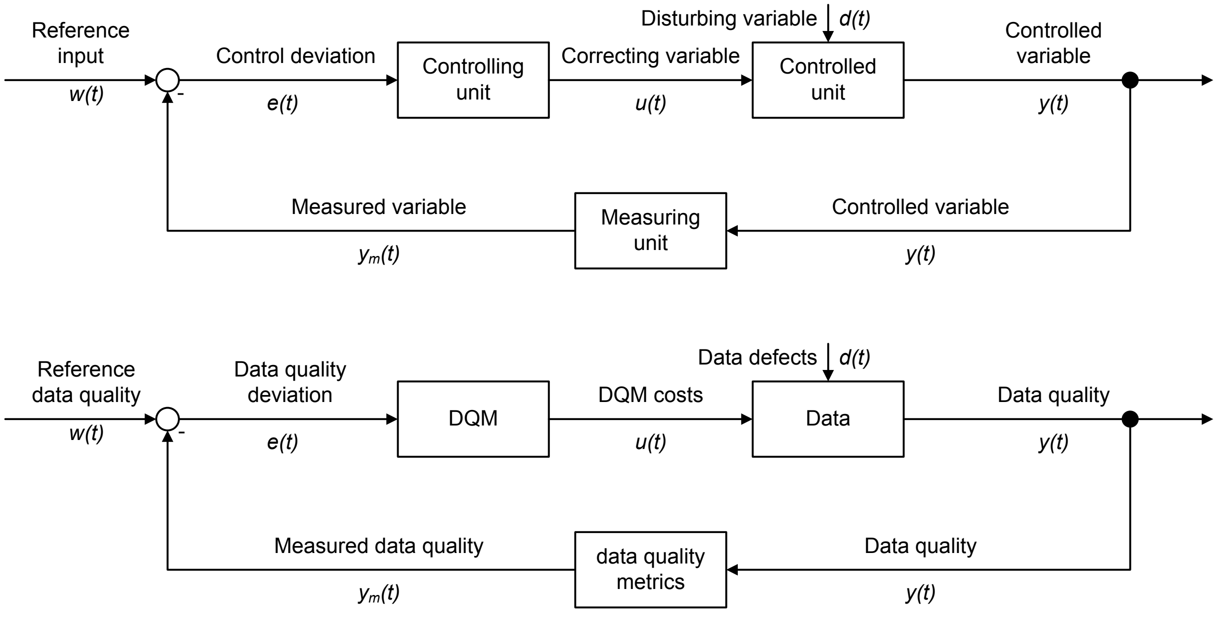

Figure 2. Block diagram of generic closed-loop control system (above) and adaptation of system elements for DQM (below)

If DQM is considered that way (i.e. as a closed-loop control system, see Table 1 for further explanation), the organizational unit responsible for DQM in a company must decide what measures for improving data quality need to be applied. To do so, expected costs of DQM initiatives planned, expected effects of these measures, expected risks of the occurrence of data defects, and expected costs caused by data defects need to be estimated. To facilitate such estimations, the paper presents a meta-model and a simulation technique, with the simulation technique being capable of calculating the data quality to be expected based on the assumptions specified in the meta-model.

Element	Generic closed-loop control system	Adaptation for DQM
Controlled unit	Dynamic system determined by the correcting variable (which can be manipulated), interfered by the disturbing variable (manipulated), and producing the controlled variable (measured).	Corporate data. The aim is to keep the data quality at a certain level.
Disturbing variable *d*	Parameter interfering with the controlled unit. Cannot be manipulated.	Data defects (see below).
Controlled variable *y*	Parameter measuring the output of the Controlled unit. Its value is dependent on the correcting variable and the disturbing Variable.	Data quality.
Measuring unit	(Dynamic) System measuring the controlled variable or – in case the controlled variable cannot be measured – calculating the controlled variable from other measured variables.	Data quality metrics (see below).
Measured variable y_m	Parameter representing the measured output of the controlled unit. Due to potential dynamics of the measuring unit itself, controlled variable and measured variable have to be distinguished.	Measured data quality. Values calculated from data quality metrics.
Reference input variable *w*	Parameter for controlling the control unit. The closed-loop control system tries to adjust the controlled variable according to this parameter and to compensate the interference of the disturbing variable. The objective is $y(t) = w(t)$ at any time.	Data quality target value.
Control deviation *e*	Difference between reference input variable and controlled variable / measured variable.	Difference between data quality target value and actual value of data quality.
Controlling unit	System controlling the controlled variable (by means of the reference input variable) so that the Controlled unit works as desired.	DQM comprising preventive and reactive initiatives.
Correcting variable *u*	Parameter by which the controlling unit tries to compensate the effect of the disturbing variable and match the controlled variable with the reference input variable.	Costs of (preventive and/or reactive) DQM initiatives.

Table 1. Elements of generic closed-loop control system adapted for DQM

Simulation Objectives

The overall aim of the DQM simulation model proposed is to explicate assumptions regarding various risks and effects (e.g. occurrence of a certain data defect or impact of a certain DQM measure) and support identification of ‘realistic’ causal relations for specific scenarios (e.g. false negative responses of a product availability check service due to data inconsistencies). A simulation model comprises those assumptions (e.g. ‘When entering customer address data into the CRM system, the probability of entering the wrong street name is 60 percent.’). An entire DQM simulation process consists of one or several simulation runs, all of which use the same simulation model (i.e. the same assumptions). A simulation run, in turn, consists of several simulation steps. In each simulation step the value of a certain variable of the simulation model is calculated (e.g. the value of a data quality metric). What is actually generated here are values of random variables, with the distribution of those random variables representing the assumptions made. Figure 3 shows the meta-model of the DQM simulation technique presented, i.e. the elements a simulation model may consist of. Denominators and semantics of the meta-model are based on industry standards (IEEE, 1998; ISO/IEC, 2007), and previous contributions to DQM research (Caballero, Verbo, Calero and Piattini, 2007; Eppler et al., 2004).

Design Objects for DQM Simulation Models

- *Business problem*. State (e.g. delivery of goods not possible) or incident (e.g. production of scrap parts) leading to decreased *process* performance and hence to poorer values of process metrics. A business problem poses a risk (in terms of probability of occurrence and intensity of impact, both of which are represented by a random variable) to a business process.

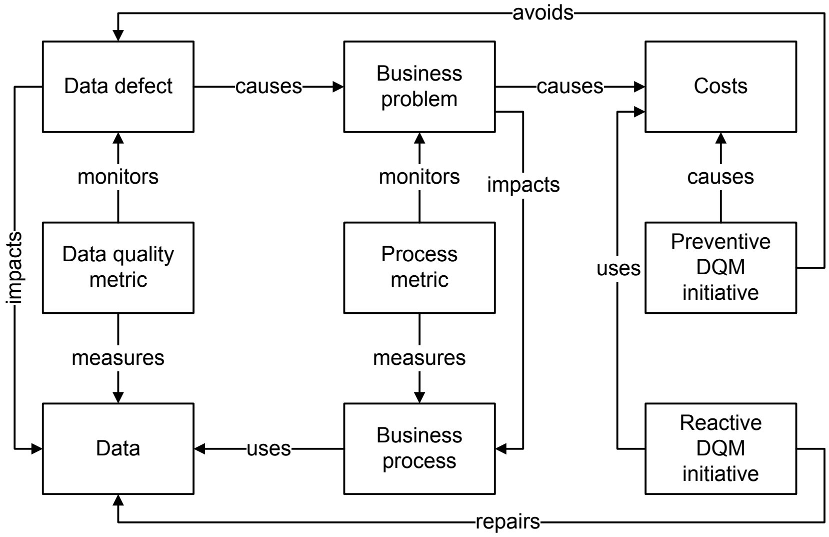

Figure 3. Design Objects for DQM Simulation Models

- *Business process*. Sequence of chronologically and typologically linked activities intended to generate a clearly defined output *bringing* about customer benefit. Transforms a given input (e.g. a certain material) into a desired output (e.g. a watch movement) under consideration of certain rules and by using certain resources (e.g. data). Is controlled and designed by means of process metrics defined as part of an overall business strategy.
- *Costs*. Measure to value the use of resources (e.g. money or material and immaterial means) for a cost objects (e.g. products, projects, *human* performance). The paper considers costs for DQM initiatives and costs caused by business problems.
- *Data*. Representations of objects (e.g. customers) and relations between objects (e.g. Customer A orders Article X) based on the description of certain object characteristics. A data element (or attribute) is a component of a data model which cannot be further *subdivided* in logical terms in a given or presupposed context. The paper considers company master data (e.g. customer address data, material data, or parts lists), with a focus not on data models (data classes and data elements) but on values assigned to data elements.
- *Data defect*. Incident leading to poorer values of data quality metrics. A data defect poses a risk (in terms of probability of occurrence and intensity of impact, both of which are represented by a random variable) to data.
- *Data quality metric*. Quantitative measure (cf. Metric, as well denoted by DQ measure (Caballero et al., 2007; ISO/IEC, 2007) of the degree to which data possess a given quality attribute (e.g. completeness, timeliness). For a data quality metric, one or more measuring methods need to be provided (i.e. where the measurement is made, what data are included, what measuring device is used, and what scale is used for measuring).
- *Preventive DQM initiative*. *Initiative* reducing either the probability of occurrence of data defects or the impact of data defects.
- *Process metric*. Quantitative *measure* (cf. Metric, as well denoted by operational measures) of the degree to which a business process possesses a given quality attribute (e.g. lead time or scrap rate). Results directly from process performance, e.g. as the average lead time for processing of an order, or as the number of order cancelations per month.
- *Reactive DQM initiative*. *Initiative* for correction of defective data (e.g. through identification and correction of false address data).

Relations between Design Objects of DQM Simulation Models

- *affects (Business Problem to Business Process)*. A *business* process can be affected by a business problem. Besides data defects there are lots of other possible causes of business problems.
- *affects (Data Defect to Data Quality Metric)*. A data *defect* (e.g. missing values for attributes of customer address data) can have a negative impact on the value of a data quality metric (e.g. completeness of customer address data).
- *affects (Data Defect to Data)*. Data can be affected by data defects (e.g. missing or wrong values for customer address attributes).

- *affects (Business Problem to Process Metric)* . A business problem (e.g. delivery of goods to the wrong address) can have a negative impact on the value of a process metric (e.g. provisioning time).
- *causes (Business Problem to Costs)*. A business problem (e.g. delay in process lead time due to delivery of goods to the wrong address) can bring about costs (e.g. lower revenues due to cancellation of orders).
- *causes (Data Defect to Business Problem)*. A data defect (e.g. wrong customer address) can bring about a business problem (e.g. delivery of goods to the wrong address).
- *measures (Data Quality Metric to Data)*. A data quality metric measures the degree to which data possess a given quality attribute (cf. Data Quality Metric).
- *measures (Process Metric to Business Process)*. A process metric measures the degree to which a business process possesses a given quality attribute (cf. Process Metric).
- *prevents (Preventive DQM Initiative to Data Defect)*. A preventive DQM initiative (e.g. implementation of a company wide data architecture) can help prevent a data defect (e.g. inconsistent data transfer).
- *repairs (Reactive DQM Initiative to Data Defect)*. A reactive DQM initiative (e.g. identification and correction of wrong data) corrects a data defect, i.e. there is a certain probability (represented by a random variable) that the initiative positively affects the value of a data quality metric affected by the data defect.

Impact Patterns

Patterns for the Impact of a Data Defect on a Data Quality Metric Measurement

When a simulation model is created (i.e. when the meta-model presented in this paper is instantiated), it is not only necessary to define values for various variables (e.g. probability of occurrence of a certain data defect) but also to specify the impact for each cause-effect relation (e.g. ‘If data defect A occurs, it must exceed a certain threshold value X to cause business problem B.’). The following sections describe a number of impact patterns that are supposed to facilitate the creation of a simulation model.

A data defect has a certain probability of occurrence and a certain impact. In a simulation step, values are generated for both random variables, i.e. a value (‚0’ or ‚1’) is generated for probability of occurrence and a value is generated for impact. The scale for measuring the impact is the same scale that measures the data quality metric affected by the data defect. For example, missing values for address data attributes have an impact on a metric for completeness of address data, which is given as a percentage value. The data defect’s impact in this case would have to be indicated as a percentage value as well (e.g. as a reduction of the value of the metric by 2 percent, cf. Figure 3). Three patterns are distinguished here (the following description follows the principle: the higher the value of the data quality metric, the higher the quality of the data):

- *Binary Pattern*. When a data defect occurs in a simulation step, the data quality metric value is reduced by a fixed value, regardless of the data defect’s impact.
- *Threshold Pattern*. When a data defect occurs in a simulation step and when the data defect’s impact exceeds a certain threshold value, the data quality metric value is reduced by a fixed value.
- *Linear Correlation Pattern*. When a data defect occurs in a simulation step, the reduction of the data quality metric correlates with the data defect’s impact (i.e. the value for the data defect’s impact is multiplied with a certain factor and the result is deducted from the value of the data quality metric).

Patterns for the Impact of a Business Problem on a Process Metric Measurement

The patterns specifying the impact of business problems on process metrics are the same patterns specifying the impact of data defects on data quality metrics (see above).

Patterns for the Impact of a Data Defect on a Business Problem

A data defect can increase the probability of occurrence and/or the impact of a business problem. Four patterns are distinguished here:

- *Binary Pattern*. When a data defect occurs in a simulation step, the probability of occurrence and/or the impact of a business problem is increased by a fixed value, regardless of the data defect’s impact.
- *Threshold Pattern*. When a data defect occurs in a simulation step and when the data defect’s impact exceeds a certain threshold value, the probability of occurrence and/or the impact of a business problem is increased by a fixed value.

- *Linear Correlation Pattern*. When a data defect occurs in a simulation step, the probability of occurrence and/or the impact of a business problem is increased by a fixed value correlates with the data defect's impact (i.e. the value for the data defect's impact is multiplied with a certain factor and the result is added to respective value).
- *Indirect Pattern*. The probability of occurrence and the impact of a business problem are calculated over the value of a data quality metric (cf. Linear Correlation Pattern). Thus they are only indirectly affected by data defects. If need by the indirect dependency can also be modeled considering a threshold (cf. Threshold Pattern).

Patterns for the Impact of a Business Problem on Costs

The patterns specifying the impact of business problems on costs are the same patterns specifying the impact of data defects on business problems (see above), with the costs' impacts simply being their monetary value.

Patterns for the Impact of a Preventive DQM Initiative

A preventive DQM initiative (e.g. implementation of a company wide data architecture) can help prevent a data defect by reducing its risk. The effect of a preventive DQM initiative (i.e. the risk reduction given by new parameters for the probability distribution) has a duration (i.e. the simulation procedure uses the lower risk for a given number of simulation steps). Two patterns are distinguished here:

- *Binary Pattern*. The initiative changes the risk parameters (i.e. the parameters of the random variables for probability of occurrence and impact) by a fixed value.
- *Initiative Booster*. The initiative reinforces the impact of another preventive or reactive DQM initiative (cf. Binary Pattern). An example would be the definition and operationalization of a DQM strategy, which indirectly contributes to risk reduction, e.g. by guaranteeing sustained financing for DQM for consistent maintenance of a data architecture.

Simulation Procedure

```
dataQ    = initial data quality
processQ = initial process quality
costs    = 0

FOR step = 0 TO duration
    FOR EACH initiative IN preventiveInitiatives
        IF initiative.IsActive()
            FOR EACH defect IN listOfDataDefects
                initiative.AdjustDataDefect(defect)
            END FOR
            IF initiative.GetStart() == step
                costs += initiative.GetCosts()
            END IF
        END IF
    END FOR
    FOR EACH defect IN listOfDataDefects
        DataQ -= defect.GetImpact()
    END FOR
    FOR EACH initiative IN reactiveInitiatives
        IF initiative.IsActive(step)
            dq += initiative.GetImpact(step)
            costs += initiative.GetCosts()
        END IF
    END FOR
    FOR EACH problem IN listOfBusinessProblems
        FOR EACH defect IN listOfDataDefects
            IF problem.IsCausedBy(defect)
                pq -= problem.GetImpact(defect)
            END IF
        END FOR
        costs += problem.GetCosts()
    END FOR
END FOR
```

Listing 1: Procedure of a DQM Simulation Run

Listing 1 describes the simulation procedure in a pseudo code notation in order to explain how the values for data quality, process quality and costs are calculated. The DQM simulation technique only considers scenarios involving one data quality metric (e.g. consistency of address data) and one process metric (e.g. provisioning time). Each simulation step calculates

- the impact of preventive DQM initiatives,
- the probability of occurrence and the impact of data defects,
- the impact of reactive DQM initiatives,
- the value of the data quality metric,
- the probability of occurrence and the impact of business problems,
- the value of the process metric,
- the cumulated costs caused by DQM initiatives and business problems.

Procedure Model for Application

The steps listed in the following constitute a procedure model that can be followed to create a simulation model for simulating scenarios.

- *Simulation Scale Specification*. To specify measuring units for the simulation steps. This value defines a simulation run's temporal interpretation. For example, in a simulation run comprising 40 steps the development of DQ may be simulated over a period of 40 days, 40 weeks, or 40 months.
- *Metric Specification*. To specify measuring units for the data quality metric (e.g. 'percent' for a metric measuring data consistency), the process metric (e.g. 'days' for a metric measuring provisioning time), and the respective initial values (e.g. '8' as average provisioning time without any interference by business problems). If a simulation refers to costs, a currency must be specified as well.
- *Data Defect Specification*. To specify data defects by means of two random variables, probability of occurrence and impact.
- *Business Problem Specification*. To specify business problems by means of two random variables, probability of occurrence and impact.
- *Metric Impact Specification*. To specify the impact of data defects on data quality metrics and the impact of business problems on process metrics, respectively (see impact patterns).
- *Problem Impact Specification*. To specify the impact of data defects on business problems and the impact of business problems on costs (see impact patterns). It must be decided for each business problem whether it is considered as independent from the data defects modeled or whether it is affected by the data defects modeled (and if so, how).
- *Initiative Specification*. To specify preventive and reactive DQM initiatives. For each initiative a period of time (starting point and duration, cf. function *IsActive* in Listing 1) and the costs expected must be indicated. The initiative's impact can be described as a random variable too, in order to simulate its probable impact.
- *Simulation*. To simulate a modeled scenario. The simulation procedure calculates the values to be simulated (e.g. metric values or costs). From the results of several simulation runs expected values (e.g. for the data quality metric) and statistical variances can be calculated.

EXAMPLE OF APPLICATION

Business Scenario

To evaluate the meta-model and the simulation technique a simulation model was developed for and applied at an international telecommunications company (employing approx. 250,000 people and having a turnover of 60 billion Euro in 2009). The scenario modeled refers to the process of customers ordering IP-TV. Part of this process is a validation service offered by the company for the customer to see whether IP-TV is available at their residence. The service's response is either positive confirming availability of the product (approx. 80 percent of responses) or negative (six different responses possible; e.g. 'Address unknown' or 'Resources for production of IP-TV used up').

Samples taken from this validation service have revealed that as much as five percent of the negative responses are false, i.e. IP-TV actually would be available at these residences. With an average of 5,000 requests per day concerning product availability and 20 percent of these requests resulting in signed contracts the company has to face lost revenues of approx. 9 million Euro per year, given an estimated value of each IP-TV customer of approx. 500 Euro per year.

As the reason for false negative responses inconsistencies between the data used for the product availability check and the data stored in the source systems (e.g. network infrastructure data, contractual data) could be identified. These inconsistencies were mapped onto rules for validation, which constitute the techniques for measuring six data quality metrics, with each data quality metric referring to the possible causes of one of the six false negative responses. The rules for validation are pairwise disjoint, i.e. a cause of a false negative response validated by one rule is not validated by another rule.

Simulation Model and Application

Based on the measuring of data quality metrics the risk of the occurrence of a specific data defect (i.e. the probability of a validation rule identifying a data defect) can be quantified. Table 2 shows the parameters of the pertinent, normally distributed random variable. Also, the table shows which data defect (DD) is used for measuring which data quality metric (indicated as errors, E). Figure 4 depicts the simulation (based on the parameters for Sim1 given in Table 2) of data defects over a period of 24 months, while Figure 5 depicts the simulation of false negative responses over the same period of time.

Data defect	DD1	DD2	DD3	DD4	DD5	DD6	DD7	DD8	DD9
Risk Sim1	953, 61	734, 50	1146, 127	1028, 37	428, 35	643, 71	495, 84	761, 129	369, 62
Risk Sim2	477, 31	367, 25	573, 64	514, 19	214, 18	322, 36	248, 42	381, 65	185, 31
Caused error	E1	E1	E2	E3	E3	E4	E5	E5	E6

Table 2. Parameters (i.e. mean, variance) of normal distributed risk of data defects (occurrence probability = 1 for all defects)

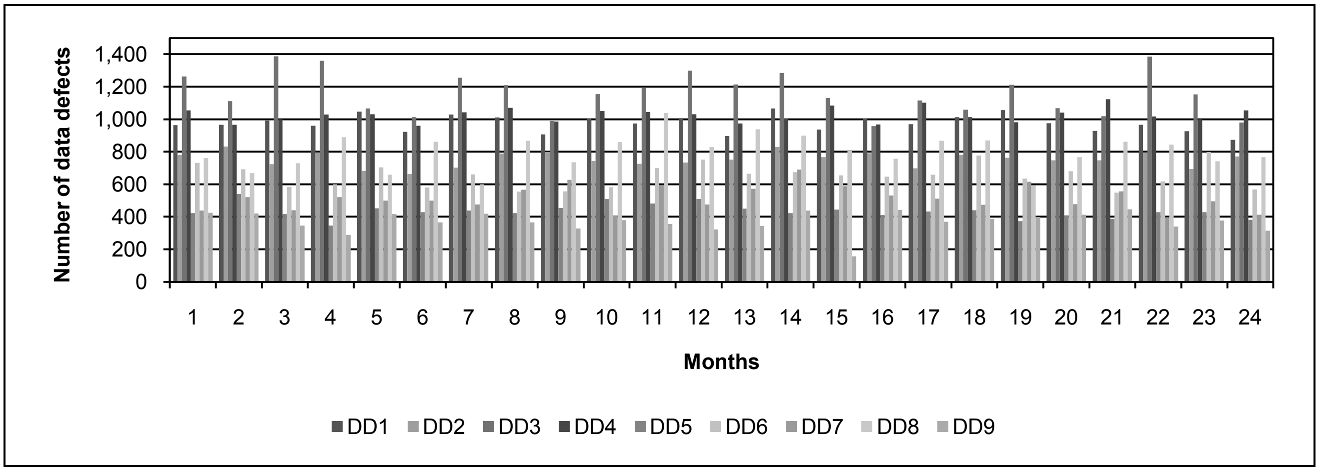

Figure 4. Simulation of data defects over a period of 24 months

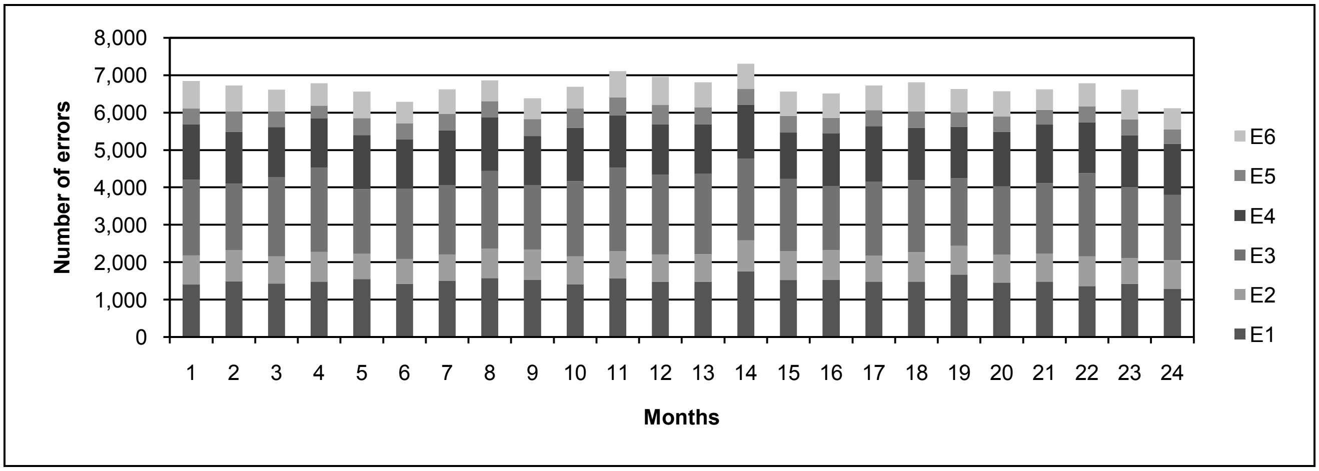

Figure 5. Simulation of false negative responses over a period of 24 months

Due to the already mentioned ratio of product availability requests and signed contracts the monthly value of a false negative response equals 20 percent (= 8,33 Euro) of the monthly value of a customer (= 41,67 Euro). Figure 6 depicts the cumulated costs (i.e. the total of lost revenues due to false negative responses of the product availability check) incurred over the entire period of time for Sim1. Furthermore, Figure 6 shows the result of a simulation run (Sim2) calculating the effect of a preventive DQM initiative (i.e. an improved data architecture). The architecture's design and implementation cost 500,000 Euro (projected on the first month of the simulation) but it also avoids 50 percent of the risk of data defects (see Table 2, second row). Therefore, the cumulated costs are lower for the preventive DQM scenario.

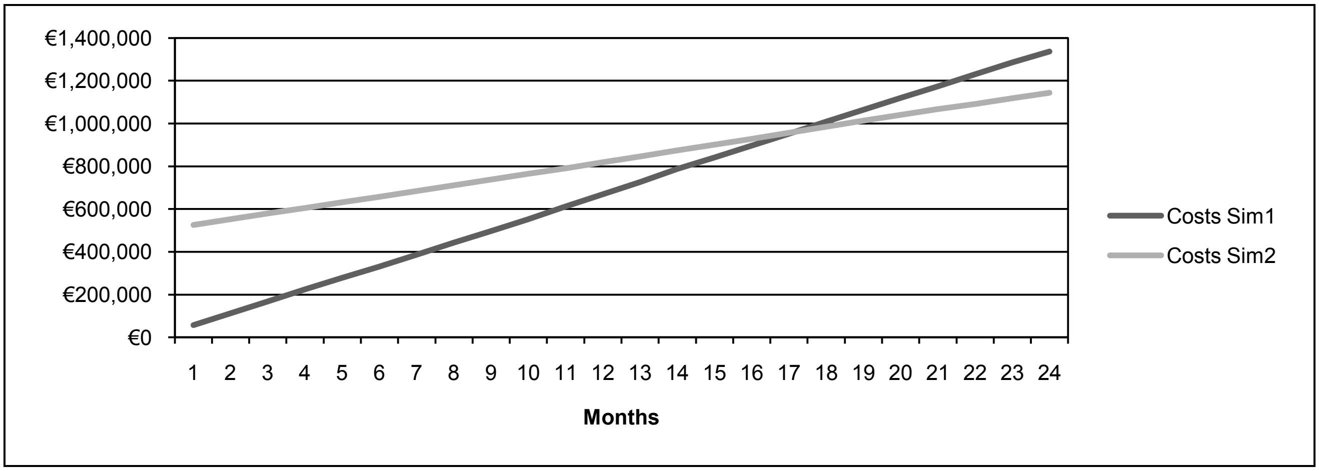

Figure 6. Cumulated costs due to false negative responses, on the right with data defect risk reduction by preventive DQM

DISCUSSION AND FUTURE WORK

The paper presents a cybernetic view on data quality management (DQM), with data quality being the controlled variable and costs of DQM initiatives being the correcting variable in a closed-loop control system. Also, the paper introduces a meta-model for description of DQM scenarios by means of a simulation model supposed to help an organizational unit responsible for DQM select appropriate measures. To facilitate scenario modeling the paper specifies various impact patterns representing relations between various elements of the model plus a simulation model. Both the meta-model and the simulation technique are then evaluated in a business scenario at an international telecommunications company.

When using the simulation model the main challenge is to identify and quantify causal relations. While this process can be delineated by a procedure model, in real-world cases the information required (e.g. frequency of the occurrence of a specific data defect) has to be collected from various sources within a company (e.g. by means of interviews and expert assessments) and mapped onto the elements of the model (e.g. the parameters for distribution of a data defect). This requires both availability and willingness to cooperate on the part of subject matter experts and sufficient knowledge about the simulation model.

Need for further research is seen in further validating appropriate simulation models and in comparing the simulated development of data quality with real effects. Although a simulation model can take into consideration only selected aspects of reality, the aim is to describe real developments at least by approximation. To be able to do so, real measuring values of data quality metrics and process metrics must be available over a substantial period of time, and the effect of DQM initiatives on such measuring values needs to be traceable. Furthermore, research should focus on further specification of the meta-model, particularly with regard to specifying different types of costs and benefits as well as different cost bearers and beneficiaries. Through such further specification DQM simulations would not just be able to explicate the basic benefit of DQM initiatives, but it would be possible also to identify those company departments likely to expect the biggest benefit from a specific DQM measure, which could then bear a bigger proportion of the costs. Finally, further research refers to the use of the simulation model for a detailed determination of the behavior of DQM as a system. Simulation results and knowledge about companies' individual preferences would allow to select the controlling unit most appropriate (e.g. PID control) including suitable parameters.

REFERENCES

1. Amram, M. and Kulatilaka, N. (1998) Real Options: Managing Strategic Investment in an Uncertain World HBS Press, Boston.
2. Batini, C., Barone, D., Mastrella, M., Maurino, A. and Ruffini, C. (2007) A Framework and a Methodology for Data Quality Assessment and Monitoring, in M.A. Robbert, R. O'Hare, M.L. Markus and B. Klein (Eds.) *Proceedings of the 12th International Conference on Information Quality*, Cambridge, 333-346.
3. Black, F. and Scholes, M. (1973) The Pricing of Options and Corporate Liabilities, *Journal of Political Economy*, 81, 3, 637-654.

4. Burgess, M. S. E., Gray, W. A. and Fiddian, N. J. (2004) Quality Measures and the Information Consumer, in S. Chengalur-Smith, L. Raschid, J. Long and C. Seko (Eds.) *Proceedings of the 9th International Conference on Information Quality*, Boston, 373-388.
5. Caballero, I., Calero, C., Piattini, M. and Verbo, E. (2008) MMPro: A Methodology based on ISO/IEC 15939 to Draw up Data Quality Measurement Processes, in P. Neely, L. Pipino and J.P. Slone (Eds.) *Proceedings of the 13th International Conference on Information Quality*, Cambridge.
6. Caballero, I., Verbo, E., Calero, C. and Piattini, M. (2007) A Data Quality Measurement Information Model based on ISO/IEC 15939, in M.A. Robbert, R. O'Hare, M.L. Markus and B. Klein (Eds.) *Proceedings of the 12th International Conference on Information Quality*, Cambridge, 393-408.
7. English, L. P. (1999) Improving Data Warehouse and Business Information Quality Wiley, New York.
8. Eppler, M. J. and Helfert, M. (2004) A Classification and Analysis of Data Quality Costs, in S. Chengalur-Smith, J. Long, L. Raschid and C. Seko (Eds.) *Proceedings of the 9th International Conference on Information Quality*, Cambridge, 311-325.
9. Fisher, C. W. and Kingma, B. R. (2001) Criticality of data quality as exemplified in two disasters, *Information & Management*, 39, 2, 109-116.
10. Gebauer, M., Caspers, P. and Weigel, N. (2005) Reproducible Measurement of Data Field Quality, in F. Naumann, M. Gertz and S. Madnick (Eds.) *Proceedings of the 10th International Conference on Information Quality*, Cambridge.
11. Gustavsson, M. (2006) Information Quality Measurement. Considerations when defining Measures supporting MPC decision making, in J. Talburt, E. Pierce, N. Wu and T. Campbell (Eds.) *Proceedings of the 11th International Conference on Information Quality*, Cambridge.
12. Heinrich, B., Klier, M. and Kaiser, M. (2009) A Procedure to Develop Metrics for Currency and its Application in CRM, *Journal of Data and Information Quality*, 1, 1, 1-28.
13. Hevner, A. R., March, S. T., Park, J. and Ram, S. (2004) Design Science in Information Systems Research, *Management Information Systems Quarterly*, 28, 1, 75-105.
14. IEEE (1998) IEEE Standard for a Software Quality Metrics Methodology, IEEE Std 1061-1998, The Institute of Electrical and Electronics Engineers.
15. ISO/IEC (2007) ISO/IEC 15939. Systems and software engineering – Measurement process, ISO/IEC 15939:2007(E), International Organization for Standardization.
16. Joshi, K. and Rai, A. (2000) Impact of the quality of information products on information system users' job satisfaction: an empirical investigation, *Information Systems Journal*, 10, 4, 323-345.
17. Lee, Y. W., Pipino, L. L., Funk, J. D. and Wang, R. Y. (2006) Journey to Data Quality MIT Press, Boston.
18. Lunze, J. (2008) Regelungstechnik 1. Systemtheoretische Grundlagen, Analyse und Entwurf einschleifiger Regelungen, (7 ed.) Springer, Berlin.
19. Naumann, F. and Rolker, C. (2000) Assessment Methods for Information Quality Criteria, in B.D. Klein and D.F. Rossin (Eds.) *Proceedings of the 5th Conference on Information Quality*, Cambridge, 148-162.
20. Orr, K. (1998) Data Quality and Systems Theory, *Communications of the ACM*, 41, 2, 66-71.
21. Otto, B., Hüner, K. M. and Österle, H. (2009) Identification of Business Oriented Data Quality Metrics, in P. Bowen, A.K. Elmagarmid, H. Österle and K.-U. Sattler (Eds.) *Proceedings of the 14th International Conference on Information Quality*, Potsdam, 122-134.
22. Peffers, K., Tuunanen, T., Rothenberger, M. A. and Chatterjee, S. (2008) A Design Science Research Methodology for Information Systems Research, *Journal of Management Information Systems*, 24, 3, 45–77.
23. Price, R. and Shanks, G. (2005) A semiotic information quality framework: development and comparative analysis, *Journal of Information Technology*, 20, 2, 88-102.
24. Redman, T. C. (1996) Data Quality for the Information Age Artech House, Boston.
25. Wand, Y. and Wang, R. Y. (1996) Anchoring Data Quality Dimensions in Ontological Foundations, *Communications of the ACM*, 39, 11, 86-95.
26. Wang, R. Y. and Strong, D. M. (1996) Beyond Accuracy: What Data Quality Means to Data Consumers, *Journal of Management Information Systems*, 12, 4, 5-34.