Measuring Master Data Quality

Findings from an Expert Survey

Boris Otto, Verena Ebner

Institute of Information Management, University of St. Gallen

1 Introduction

1.1 Motivation

Data quality management (DQM) plays a critical role in all kinds of organizations (Pipino et al. 2002). With data being the foundation for information it is one of the most important criteria for strategic business decisions within organizations. That is why poor data quality can have a critical effect on business processes leading to increased costs and lowering customer and employee satisfaction (Redman 1998).

Nearly every IT system has "dirty" (erroneous) data. About 75 per cent of organizations have identified costs originating from dirty data (Marsh 2005). U.S. businesses pay $600 billion a year due to a lack of data quality (Eckerson 2002).

To improve data quality (DQ) as well as to evaluate the current status, the improvement of data quality and the effect of data quality initiatives have to be measured. Several authors point out that: "Only what can be measured can be improved" (Wand & Wang 1996, Wang & Strong 1996, English 1999). What is needed is a measurement approach to determine the level of data quality over time.

Many companies are running data quality initiatives (White et al. 2006) focusing on actions to increase data quality. But how many of those really measure the quality of their data and how many really know how to implement a measurement system? The research presented in this paper was carried out in the context of the Competence Center Corporate Data Quality (CC CDQ), a research program on master data quality at the Institute of Information Management at the University of St. Gallen.

1.2 Research Question and Paper Structure

The research question addressed in this paper focuses on the progress of organizations regarding the measurement of data quality. That means: "do organizations

already measure data quality?", "what do they measure?" and "how do they measure?". The research method used is a market survey with data quality experts as interview partners.

The research paper presents an overview of the industries status in measuring data quality and gives an indication of the dissemination among practitioners and their understanding of the task at hand. This study does not include a detailed description of the requirements and actions necessary to implement a measurement. The requested tasks do not guarantee completeness.

The remainder of the paper is structured as follows. The following chapter gives an overview of the background information, especially the terminology of data, data quality and the progress of research in the area of measuring data quality. Section 3 describes the research approach and the underlying framework which was applied in the study. In section 4 the essential results of the survey and the result evaluation are presented. The paper closes with a summary and an outlook on further research objectives and activities planned.

2 Background

2.1 Master Data and Master Data Management

Information systems provide data in a certain business context. When data is being used by human beings, they turn into information, and information finally turns into knowledge by interpreting and linking information for a given purpose (Davenport & Prusak 1998, Spiegler 2000, Boisot & Canals 2004, Stahlknecht & Hasenkamp 2005).

Master data stores and describes features of a company's core entities, which most notably are customers, suppliers, products, materials, and employees (DAMA 2008, Dreibelbis et al. 2008, Loshin 2008). Typically, master data is used across multiple business processes (e. g. supplier master data is used both by procurement departments and by accounting departments) and is often stored in and/or used by multiple application systems.

Master data can be distinguished from other data, such as transaction data or inventory data, using the following criteria (Mertens 2000, White et al. 2006, Dreibelbis et al. 2008): time reference, modification frequency, volume stability and existential independence.

The master data regarded in the organizations varies between industries as well as between organizations. But there are commonalities between the core classes on the basis of which a classification can be applied (Loshin 2008). Dreibelbis et al (2008) identify three categories: product, party and account addressing the questions "Who?", "What?" and "How?". The domain location is of relevance for all of the classes, so it can be seen as a subdomain of master data. For each domain they

provide the most important classes. The selected classes with their corresponding categorization are shown in Table 1.

Table 1: Master data domains.(Dreibelbis et al. 2008)

Who? (party)	What? (product)	How? (account)
customers	products	accounts
suppliers	materials	contracts
employees	assets	
organizations	services	

Master data management (MDM) aims at creating an unambiguous understanding of a company's core entities (Smith & McKeen 2003). As an application-independent process, it ensures the consistency and accuracy of these data by providing a consistent understanding and trust of master data entities supported by mechanisms to use consistent master data throughout the organization and managing the change. These goals are achieved by implementing a corporate framework including the domains organization, processes and architecture (Dreibelbis et al. 2008).

2.2 (Master) Data Quality and Data Quality Management

Data is defined of high quality if it has the ability to satisfy the requirements for its intended use in a specific situation. This is often referred to as "fitness for use" (English 1999, Redman 2001, Olson 2003). The intended use is commonly described as a multi-dimensional construct consisting of a set of quality attributes, called data quality dimensions which are determined by the data consumer (Wang & Strong 1996). On the one hand, DQ plays an important role in the success of a MDM. It supports the trustworthiness of master data and its techniques can be applied in the implementation process. On the other hand, MDM can improve DQ. It reduces the error rate as there is a high probability of identifying mistakes by the integration of multiple systems (Loshin 2008).

Within the study we used six of the major dimensions proposed by Wang (1996), English (1999) and Redman (1996) (c.f. Table 2).

Table 2: Data Quality Dimensions

Quality Dimension	Definition
accuracy	The extent to which data are correctly representing an action or real world object.
completeness	The extent to which values are present in a data collection
timeliness	The extent to which data represents the real world at a given point in time.
consistency	The extent to which data knowable in one database correspond to the data in a redundant or distributed database.
relevancy	The extent to which data is applicable and helpful for the task at hand.
accessibility	The extent to which data is available at a given point in time.

3 Research Approach

3.1 Conceptual Framework

Considering data as a product, a distinction between the product with its requirements and the production process with its actors can be made (Wang 1998). Looking further, the actions of the production process take place in the business processes lying in between the production process and the data products and interconnecting them. Upon this distinction we identified three points of measurement: the product, i.e. the data stored within the data base, the production process, in the following named as data lifecycle process, and the connection between these elements, referred to as business processes.

The measurement characteristics applied within the product-part correspond to the data quality dimensions as defined in chapter 2.2. For the production process, the key performance indicators (KPIs) used include cycle time, idle time (e.g. between workflow steps), status of lifecycle (e.g. completed, approved, deactivated), process costing and adherence to the process. The relationship between these elements depicting the impact of the product onto the business processes can be measured as loss of quality, time and costs. In the following the product-part is referenced to as "measuring at the data base", the production process as "data lifecycle" and the element connecting the product and production process as "impact on business processes".

3.2 Sampling, data collection and analysis

The purpose of the study is descriptive (Cavana et al. 2001). It aims at gaining an overview on the progress of companies in measuring data quality.

For data collection an online questionnaire was used. The survey covered 27 questions grouped in five parts with 2-9 questions each. The average response time for completing the questionnaire was approximately 25 minutes. The survey was open for two weeks during June 2009 and the participants were invited via e-mail. The first block contains demographic information. The next three sections address the building blocks of the conceptual framework. The last one treats aggregated questions addressing challenges and problems within measurement projects. The complete questionnaire is available at http://tr.im/dq_survey.

The questions were basically semi-open, i.e. the answers were predefined with the possibility to give a specification, optional and multiple-choice. For this reason a response rate of over 100 per cent as well as an inconsistent response rate is possible. The intent was to give respondents the possibility to skip questions if there is no measurement implemented. Open questions were used where it was intended to describe a procedure or the like.

The sample consists of approximately 300 members of an e-mail distribution list. The list comprises members of the CC CDQ project network. This network has been used to exchange experiences on the subject matter since three years. All contacts have roles related to MDM or DQ, like (Master) Data Officer and (Master) Data Manger and are able to make a statement about the company's improvement on data quality management. The experts all belong to large organizations headquartered in Germany, Switzerland and Austria. The industrial sectors[1] are manufacturing (>50%), transportation & communication, financial & insurance activities, construction, electricity, gas and water supply as well as wholesale and retail trade.

The sample size and selection does not support statistical representativeness or exploratory results. Since the nature of the study is descriptive, however, the research design is adequate and in line with demands for research pragmatism (Strübing 2008).

4 Result Presentation and Interpretation

In total, 41 of the nearly 300 experts completed the survey which results in a response rate of about 15 percent. In the subsequent sections the results are presented followed by an interpretation.

[1] corresponding to the Statistical Classification of Economic Activities in the European Community, Rev. 2 (NACE Rev. 2)

4.1 Measuring master data quality at the data base

In the section "measuring at the data base" the master data classes measured most are customers (27), materials (24), products (23) and suppliers (18). Table 3 shows that the dimensions measured most are consistency, completeness and timeliness.

Table 3: Data quality dimensions measured

Data quality dimension	Number of responses
accuracy	13
completeness	18
timeliness	16
consistency	20
relevancy	8
availability	12

The point of the measurement within the process or system is in 76 % within a data base (system), 73% are measuring during the maintenance of data. During data-import and -export about 62% respondents measure and 32% during data exchange (e.g. within middleware).

With regard to the frequency of measurement most measure "ad hoc/occasionally" (58%) and "monthly" (56%). A continuous measurement, for example on each data import is used by 39%. There are 28% with an occurrence less then monthly and 11.2% more than monthly.

The main responsible for DQ measurement are the data (59%) and process owner (53%) and some "other responsible person" (29%) very often stated as a data quality management team. Less mentioned is the system owner (24%) and just 6% have an accountable, external service partner.

For displaying the values, a presentation in comparison to a target value (74%) and in contrast to earlier measurement values (68%) is usual. The usage of values of a corresponding group (e.g. locations, systems, processes, industries) is less common (38%).

The results of the measurement at the data base are normally reported to the process (61%) and data owner (56%). The executive board is at least in 31% of the cases the recipient and the system owner in 25%.

On the question if data quality is integrated in Service Level Agreements only 23% of the respondents answered with "yes". These agreements were specified with: contracts with service provider or personal target agreements. One respondent described checks on actuality, completeness, consistency, correctness, trans-

parency and uniqueness, another mentioned given target values as part of the agreements.

4.2 Measuring master data lifecycle management

As you can see in Table 4 the most important KPIs for data lifecycle management are "status of lifecycle" (12), "cycle time" (11) and "idle time" (8).

Table 4: Key Performance Indicators measured

Key Performance Indicator	Number of responses
cycle time	11
idle time (e.g. between workflow steps)	8
status of lifecycle (e.g. completed, approved, deactivated)	12
process costing	1
adherence to the process	4
other	4

The responsible person for the data lifecycle measurement is at most of the organizations the data owner (61%) followed by the process owner (43%) as well as the system owner (25%) on the same level as other roles (e.g. data quality management team) (22%). External service partners are marginal.

The reporting of the values is used in comparison to values of earlier points in time in 82%, to a target value in 57.1% and to values of corresponding groups in 32% of the cases.

Similar to the measurement at the database, reports address the process owner (70%) and data owner (53%), followed by the executive board (30%) and the system owner (10%).

For the execution of the measurement as well as for the reporting several tools from various vendors exist. The measuring tools used the most are Microsoft tools like MS Access and MS Excel (7 respondents). Software from SAP (4 respondents) and IBM (3 respondents) were also mentioned as well as MioSoft (1), Fuzzy Post / Fuzzy Double (1), Informatica (1) and Geschäftslogic (1). Tools for reporting used by the participants are MS Office (17), Business Warehouse Reports (2), Business Objects (1) and Oracle Application Express (APEX) (1).

4.3 Measuring the business impact of master data quality

The section "measuring the impact on business processes" was returned by a comparatively small number of respondents (10).

In this area the value-added processes are measured regarding time, costs and quality. The results in this domain show that most of the companies give their attention to the process quality which is measured by 7 of 10 interviewees. 3 are measuring losses in time and the same number measures increased process costs. The processes used are mainly marketing & sales (7) and service (6). Operations (4), inbound (3) and outbound logistics (3) are subsequent.

Measuring the adherence to business rules[2] is already established in 8 out of 14 organizations. The execution of the validation was stated as follows: internal and external audits, deviation from a threshold, consistency checks among systems and dependencies within a data set (sanity checks), tool-based validation of data quality requirements, standard rules described in the process documentation (rules for business processes) and input validation.

The responsibility for the measurement of business processes rests in most of the cases with the process owner (8) followed by the data owner (6). A data quality expert is mentioned 3 times, the system owner as well as an external service provider is named twice.

Within reports, mostly a representation of the value's sequence in time (10) and the comparison to a threshold (9) is chosen (10). The opposition of the measurement with a differing measuring is rarely used (4).

The reported receiver is normally the process owner (8) as well as the data owner (8). In 6 cases the addressee is the executive board and in 3 responses the system owner.

Data quality KPIs are integrated in the target agreement in 4 organizations. The implementation is realized by data quality target values or a data quality index.

4.4 Questions on challenges and success factors

The last section of the survey contained general questions covering the estimation of the importance of data quality measurement and a statement on challenges and success factors. Of the given three areas the measurement at the data base was assessed as the most important (79% - very important). The succeeding section is the impact on business processes. Measuring data lifecycle management is stated as least important.

Gathering the challenges and problems within the area of measuring data quality the statements were: determination of data quality rules, data aggrega-

2 "A business rule is a statement that defines or constrains some aspect of a business. The intent of a business rule is to control or influence some aspect of a business through the imposition of structure." Ross, RG (2003) *Principles of the Business Rule Approach.* Addison-Wesley Information Technology.

tion/reduction of complexity, determination of the impact on business rules, consistency and repeatability of the approach, awareness, motivation and target value definition.

The success factors mentioned are: defined standards, awareness and management support, holistic approach, individual target agreement, clearly defined responsibilities and processes, visualization of costs and effects, simplicity and centralized governance.

4.5 Result discussion

As a general outcome, the survey shows that most of the organizations taking part in the study measure data quality at the data base. Data lifecycle management comes second and the impact on business processes ranks third. In the course of the questionnaire, a clear trend towards a decreasing answering rate can be observed. On the one hand this points at low measuring activities in these sections. On the other hand it points at a missing understanding or differentiation of the parts.

Between the measured master data classes of the data base measurement and the data lifecycle management a correlation can be detected. Most of the participants, measuring data quality dimensions at certain master data classes are measuring KPIs in nearly all of these master data classes. There are a few cases in which there are KPIs measured also in other data classes.

The responsible person of the measurement is in all sections primarily the data owner, except in the area covering the impact on business processes. Here the process owner is most often responsible for the measurement of process delays regarding time, costs and quality. Within the section data lifecycle, the process and data owner tend to be identical.

The presentation of the reports is differing in the three parts. In the data base section the measurement of different values makes a comparison of those values possible. The KPIs as well as the impact on costs, time and quality is often not comparable. In these sections the values are presented in contrast to a target value or as time series.

The reporting is also comparable among the sections. The reporting is mainly addressed at the process or data owner. But further, a lot of respondents report to the executive board, especially in the area of measuring the impact on business processes.

5 Summary and Outlook

5.1 Major findings

Although only 40 experts answered the survey, the study gives an overview of the status quo in measurement of data quality. It shows that there are companies already measuring data quality. It also states that there is not a clear understanding of the relevant tasks and there is a need for a further definition.

Organizations measuring data quality are mainly focusing on the measurement of the data quality metrics measuring the quality of the data within the data base. The data lifecycle KPIs and even more the impact on product quality, costs and execution time are less important to most of the interviewees. The importance of measuring data quality at the origin of its problems, where mistakes can be identified before they arise right away, has to be stated.

5.2 Outlook to future research

Against the background of the study at hand future research will focus on the development of a framework containing all tasks necessary to implement a measurement system. The framework should assess the complete extent of data quality problems from its root causes to its impacts.
After collecting specific measurement initiatives best practices will be derived. An approach for implementing data quality measurements will be developed. A sample evaluation as well as reconciliation with the target group is planned to verify the findings.

References

Boisot, MH and Canals, A (2004) Data, information and knowledge: have we got it right? Journal of Evolutionary Economics 14, 43 -67.

Cavana, RY, Delahaye, BL and Sekaran, U (2001) Applied Business Research - Qualitative and Quantitatvie Methods. John Wiley & Sons, Milton.

DAMA (Ed.) (2008) The DAMA Dictionary of Data Management. Technics Publications.

Davenport, TH and Prusak, L (1998) Working Knowledge - How Organizations Manage What They Know. Harvard Business School Press, Boston (MA).

Dreibelbis, A, Hechler, E, Milman, I, Oberhofer, M, van Run, P and Wolfson, D (2008) Enterprise Master Data Management: An SOA Approach to Managing Core Information. Pearson Education, Boston.

Eckerson, W (2002) Data Quality and the Bottom Line: Achieving Business Success through a Commitment to High Quality Data. The Data Warehousing Institute, Seattle, WA.

English, LP (1999) Improving Data Warehouse and Business Information Quality. John Wiley & Sons, Inc., New York, NY.

Loshin, D (2008) Master Data Management. Elsevier Science & Technology Books, Burlington, MA.

Marsh, R (2005) Drowning in dirty data? It's time to sink or swim: A four-stage methodology for total data quality management. Database Marketing & Customer Strategy Management 12 (2), 105-112.

Mertens, P (2000) Integrierte Informationsverarbeitung - Band 1: Administrations- und Dispositionssysteme in der Industrie. Gabler, Wiesbaden.

Olson, J (2003) Data Quality - The Accuracy Dimension. Morgan Kaufmann, San Francisco.

Pipino, LL, Lee, YW and Wang, RY (2002) Data Quality Assessment. Communications of the ACM 45 (4), 211-218.

Redman, TC (1996) Data Quality for the Information Age. Artech House, Boston, London.

Redman, TC (1998) The Impact of Poor Data Quality on the Typical Enterprise. Communications of the ACM 41 (2), 79-82.

Redman, TC (2001) Data Quality. The Field Guide. Digital Press, Boston.

Ross, RG (2003) Principles of the Business Rule Approach. Addison-Wesley Information Technology.

Smith, HA and McKeen, JD (2003) Developments in Practice VIII: Enterprise Content Management. Communications of the Association for Information Systems 11, 647-659.

Spiegler, I (2000) Knowledge management: a new idea or a recycled concept? Communications of the AIS 3, Article 14.

Stahlknecht, P and Hasenkamp, U (2005) Einführung in die Wirtschaftsinformatik. Springer, Berlin.

Strübing, J (2008) Grounded Theory: Zur sozialtheoretischen und epistemologischen Fundierung des Verfahrens der empirisch begründeten Theoriebildung. VS Verlag, Wiesbaden.

Wand, Y and Wang, RY (1996) Anchoring Data Quality Dimensions in Ontological Foundations. Communications of the ACM 39 (11), 86-95.

Wang, RY (1998) A Product Perspective on Total Data Quality Management. Communications of the ACM 41 (2), 58-65.

Wang, RY and Strong, DM (1996) Beyond Accuracy: What Data Quality Means to Data Consumers. Journal of Management Information Systems 12 (4), 5--34.

White, A, Newman, D, Logan, D and Radcliffe, J (2006) Mastering Master Data Management. Gartner, Stamford.