HOW TO MEASURE DATA QUALITY?

A METRIC BASED APPROACH

Bernd Heinrich **Marcus Kaiser**

Mathias Klier

Abstract

The growing relevance of data quality has revealed the need for adequate measurement since quantifying data quality is essential for planning quality measures in an economic manner. This paper analyzes how data quality can be quantified with respect to particular dimensions. Firstly, several requirements are stated (e.g. normalization, interpretability) for designing adequate metrics. Secondly, we analyze metrics in literature and discuss them with regard to the requirements. Thirdly, based on existing approaches new metrics for the dimensions correctness and timeliness that meet the defined requirements are designed. Finally, we evaluate our metric for timeliness in a case study: In cooperation with a major German mobile services provider, the approach was applied in campaign management to improve both success rates and profits.

Keywords: Data Quality, Data Quality Management, Data Quality Metrics

Introduction

In recent years, data quality (DQ) has gained more and more importance in theory and practice due to an extended use of data warehouse systems, management support systems (Cappiello et al. 2003) and a higher relevance of customer relationship management (CRM) as well as multichannel management (Cappiello et al. 2004b; Heinrich and Helfert 2003). This refers to the fact that – for decision makers – the benefit of data depends heavily on their completeness, correctness, and timeliness, respectively. Such properties are known as DQ dimensions (Wang et al. 1995). Many firms have problems to ensure DQ (Strong et al. 1997) and according to an earlier study by (Redman 1998) "the total cost of poor data quality" is between 8% and 12% of their revenues. Other studies indicate that 41% of the data warehouse projects fail, mainly due to insufficient DQ (Meta Group 1999). Furthermore, 67% of marketing managers think that the satisfaction of their customers suffers from poor DQ (SAS Institute 2003). These figures impressively illustrate the relevance of DQ today. The consequences of poor DQ are manifold: They range from worsening customer relationships and customer satisfaction by falsely addressing customers to insufficient decision support for managers.

The growing relevance of DQ has revealed the need for adequate measurement. Quantifying the current state of DQ (e.g. of customer data) is essential for planning DQ measures in an economic manner. In the following we discuss how metrics for selected DQ dimensions can be designed with regard to two objectives: (1) Enabling the measurement of DQ, (2) Analyzing the economic consequences of DQ measures taken (e.g. to what extent does data cleansing of customer’s address data improve their correctness and lead to higher profits?). The developed metrics were applied in cooperation with a major German mobile services provider. The aim of the project was to analyze the economic consequences of DQ measures in campaign management.

Taking into account the design guidelines defined by (Hevner et al. 2004), we consider the metrics as our artifact and organize the paper as follows: After briefly discussing the relevance of the problem in this introduction, the next section defines requirements that guide the process of searching for adequate DQ metrics. In section three, selected approaches are analyzed. The fourth section designs innovative, formally noted metrics for the DQ dimensions correctness and timeliness, and examines their contribution compared to existing approaches. For evaluating the designed metrics, a case study can be found in the fifth section: It especially illustrates how the metric for timeliness was applied within the campaign management of a mobile services provider and points out the economic benefit of using the metric. The last section sums up and critically reflects the results.

Requirements of Data Quality Metrics

In order to support an economically oriented management of DQ, metrics are needed to quantify DQ so as to answer questions like the following: Which measure improves DQ most? Which one has the best costs-benefit ratio?

Figure 1 illustrates the closed loop of an economically oriented management of DQ. This loop can be influenced via DQ measures (e.g. data cleansing measures, buying external address data etc.). Taking measures improves the current level of DQ (quantified by means of metrics). This leads to a corresponding economic benefit (e.g. enabling more effective customer contacts). Moreover, based on the level of DQ and taking into account benchmarks and thresholds, firms can decide on taking (further) measures or not. From an economic view, only those measures must be taken that are efficient with regard to costs and benefit (Campanella 1999; Feigenbaum 1991; Shank et al. 1994). E.g. given two mutually exclusive measures having equal economic benefit, it is rational to choose the one with lower costs.

Therefore, this paper aims at quantifying quality by means of metrics for particular dimensions. The identification and classification of DQ dimensions is treated from both a scientific and a practical point of view by many publications (English 1999; Eppler 2003; Lee et al. 2002; Jarke et al. 1997; Redman 1996; Wang et al. 1995). In this paper, we do not treat DQ dimensions (e.g. concise representation) that concern aspects like syntactic criteria or data formats, but deal with semantics of data values. Furthermore, we focus on the two dimensions correctness and timeliness (meant here as quantifying recency), since these dimensions have been paid less attention in scientific literature (see next section). The second reason for selecting these two dimensions is that the main problem of many domains like CRM and manufacturing information products is usually not the incompleteness of data. Instead, it is of higher relevance to keep huge sets of customer data, transaction data and contract data correct and up-to-date. Thirdly, it is important that the DQ level can be quantified to a large extent automatically, which leads to lower costs of measurement, especially in case of huge sets of data. Therefore, we focus on the dimension timeliness, because – as described below – it seems to be particularly adequate.

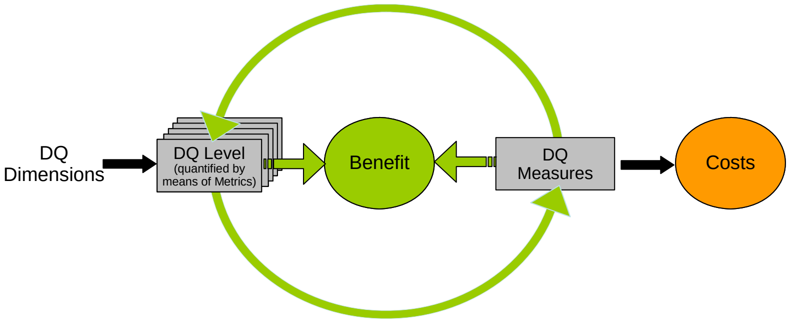

Figure 1: Data quality loop

In literature there are two different perspectives on the measurement of quality (Heinrich and Helfert 2003; Juran 2000; Teboul 1991): Quality of Design and Quality of Conformance. Quality of Design denotes the degree of correspondence between the users' requirements and the specification of the information system (e.g. specified by means of data schemata). In contrast, Quality of Conformance represents the degree of correspondence between the specification and the existing realization in information systems (e.g. data schemata vs. set of stored customer data). The distinction between Quality of Design and Quality of Conformance is important within the context of quantifying DQ: It separates the (mostly) subjective analysis of the correspondence between the users' requirements and the specified data schemata from the measurement – which is more objective – of the correspondence between the specified data schemata and the existing data values. In the following we focus on Quality of Conformance.

In practice, most DQ measures taken are developed on an ad hoc basis to solve specific problems (Pipino et al. 2002) and thus are often affected by a high level of subjectivity (Cappiello et al. 2004a). In order to enable a scientific foundation and a design evaluation of the metrics, we state the following requirements:

First, we refine the *representation consistency* by (Even and Shankaranarayanan 2007) to requirements (R 1) to (R 3):

R 1. [*Normalization*] An adequate normalization is necessary to assure that the values of the metrics are comparable (e.g. to compare different levels of DQ over time; Pipino et al. 2002). In this context, DQ metrics are often ratios with a value ranging between 0 (perfectly bad) and 1 (perfectly good) (Pipino et al. 2002; Even and Shankaranarayanan 2007).

R 2. [*Interval scale*] To support both the monitoring of how the DQ level changes over time and the economic evaluation of measures, we require the metrics to be interval scaled. This means, the difference between two levels of DQ must be meaningful. Consider for instance an identical difference of 0.2 between the values 0.7 and 0.9 and the values 0.4 and 0.6 of the metric for correctness; this means, that the quantity of data that is correct changes to the same extent in both cases.

R 3. [*Interpretability*] (Even and Shankaranarayanan 2007) demand the measurement being "easy to interpret by business users". For this reason, the DQ metrics have to be comprehensible. E.g., considering a metric for timeliness, it could be interpretable as the probability that a given attribute value within the database is still up-to-date.

(R 4) integrates the consistency principles *interpretation consistency* and *aggregation consistency* stated by (Even and Shankaranarayanan 2007).

R 4. [*Aggregation*] In case of a relational[1] data model, the metrics shall allow a flexible application. Therefore, it must be possible to quantify DQ on the level of attribute values, tupels, relations and the whole database in a

[1] If the metrics are adequate for attributes, tupels and relations, they are also usable for views as well as structured data in an Excel spreadsheet: Views are a set of attribute values and can therefore be valuated by the metrics developed according to our requirements. Moreover, (R 4) demands that the DQ of several views can be aggregated.

way, so that the values have consistent semantic interpretation (*interpretation consistency*) on each level. In addition, the metrics must allow aggregation of values on a given level to the next higher level (*aggregation consistency*). E.g., the measurement of the correctness of a relation should be computed based on the values of the correctness of the tupels being part of the relation and have the same meaning as the DQ measurement on the level of tupels.

(Even and Shankaranarayanan 2007) demand *impartial-contextual consistency* of the metrics. This refers to our requirement of the metrics being adaptive and thereby enabling a contextual perception of DQ in (R 5).

R 5. [*Adaptivity*] To quantify DQ in a goal-oriented way, it is necessary that the metrics can be adapted to the context of a particular application. If the metrics are not adapted, they should fold back to the non-adapted (impartial) measurement.

In addition, we state one more property that refers to the measurement procedure.

R 6. [*Feasibility*] For the purpose of enabling their application, the metrics are based on input parameters that are determinable. When defining metrics, measurement methods should be defined and in cases when exact measurement is not possible or cost-intensive, alternative (rigorous) methods (e.g. statistical) shall be proposed. From an economic point of view, it is also required that the measurement procedure can be accomplished at a high level of automation.

Literature Review

Literature already provides a few approaches for quantifying DQ. They differ in the DQ dimensions taken into account and in the underlying measurement procedures (Wang et al. 1995). In the following, we briefly describe some selected approaches for the DQ dimensions correctness and timeliness and analyze them with respect to the requirements above.

The AIM Quality (AIMQ) method for quantifying DQ consists of three elements (Lee et al. 2002): The first element is the product service performance model which arranges a given set of DQ dimensions in four quadrants. On the one hand, the DQ dimensions are distinguished by their measurability depending on whether the improvements can be assessed against a formal specification (e.g. completeness with regard to a database schema) or an user's requirement (e.g. interpretability). On the other hand, a distinction is made between product and service quality. Based on this model, DQ is quantified via the second element: A questionnaire for asking users about their estimation of DQ. The third element of the AIMQ method consists of two analysis techniques in order to interpret the assessments. The first technique compares the DQ of an organization to a benchmark from a best-practices organization. The second technique quantifies the distances between the assessments of different stakeholders. Beyond novel contributions the AIMQ method for DQ Measurement is based on the (subjective) users' estimation of DQ via a questionnaire. Therefore, it refers mainly to a Quality of Design definition, whereas we focus on a Quality of Conformance definition. Since the approach is not formally noted, it can hardly be analyzed whether it meets the requirements (R 1) to (R 3). It also does not deal with aggregating DQ from lower levels (e.g. correctness on the level of attribute values) to the next higher level (e.g. correctness on the level of tupels) (R 4). Moreover, they provide no possibility to adapt this measurement to a particular scope (R 5). Instead, it combines (subjective) DQ estimations of several users who generally use data for different purposes.

Besides this scientific approach two practical concepts by English and Redman shall be briefly discussed in the following. English describes the total quality data management method (English 1999) that follows the concepts of total quality management. He introduces techniques for quantifying quality of data schemata and architectures (of information systems), and quality of attribute values. Despite the fact that these techniques were applied within several projects, a general, well-founded procedure for quantifying DQ is missing. In contrast, Redman chooses a process oriented approach and combines measurement procedures for selected parts in an information flow with the concept of statistical quality control (Redman 1996). He also does not present any formally noted metrics.

In the following, we discuss two approaches (Hinrichs 2002 and Ballou et al. 1998) in detail. To our best knowledge, these are the only approaches that (1) are formally noted, (2) are based for the most part on a Quality of Conformance definition and (3) design metrics for at least one of the dimensions correctness and timeliness.

From a conceptual view, the approach by (Hinrichs 2002) is very interesting, since it aims at an objective, goal-oriented measurement. His metrics for correctness and timeliness are defined as follows.

To be determined as correct, the values of attributes in the information system must correspond to their real world counterparts. Let w_I be a value of an attribute within a database and w_R the corresponding value of the attribute in the real world. $d(w_I, w_R)$ is a domain-specific distance function quantifying the difference between w_I and w_R, normalized to the interval [0; ∞]. Examples for such distance functions are $d_1(w_I, w_R) := \begin{cases} 0 & \text{if } w_I = w_R \\ \infty & \text{else} \end{cases}$ (which is independent of the field of application), $d_2(w_I, w_R) := |w_I - w_R|$ for numeric, metrically scaled attributes and edit distance, Levenshtein distance or Hamming distance for strings. Based on such distance functions, Hinrichs defines the metric for correctness as follows:

$$Q_{Corr.}(w_I, w_R) := \frac{1}{d(w_I, w_R) + 1}$$

When applying this metric – besides others – the following problems arise: Firstly, the results are hardly interpretable (R 3). This can be illustrated by the example in Figure 2 that analyzes the correctness of the attribute “surname” and the values “Eissonhour” and “Eisenhower” as well as “Bird” and “Hunt”. The example uses the Levenshtein distance as distance function, which is given by the minimum number of operations needed to transform a string into another. Operation means here an insertion, deletion, or substitution of a single character:

$$Q_{Corr.}(\text{"Eissonhour","Eisenhower"}) = \frac{1}{d_{Lev.}(\text{"Eissonhour","Eisenhower"}) + 1} = \frac{1}{4+1} = 20.0\%$$

$$Q_{Corr.}(\text{"Bird","Hunt"}) = \frac{1}{d_{Lev.}(\text{"Bird","Hunt"}) + 1} = \frac{1}{4+1} = 20.0\%$$

Figure 2: Quantifying correctness by means of the metric designed by (Hinrichs 2002)

As the example illustrates, we get the same result (20.0%), although the surname “Eissonhour” may be identified (as the correct surname “Eisenhower”) e.g. within a mailing campaign, whereas considering “Bird” and “Hunt”, it is nearly impossible that a mailing reaches the customer (this problem arises also when using other distance functions as e.g. the Hamming distance). Four facts cause this weakness: Firstly, the range of the resulting values for correctness depends heavily on the applied distance function (R 1). Secondly, the value range [0; 1] is generally not covered, because the value only results in 0 if the number of failures is ∞. Thirdly, using a quotient leads to the fact that the values are not interval scaled (R 2), which hinders evaluation of DQ measures. Moreover, the results of the metric are not interpretable (R 3) due to the quotient, i.e. the metric is hardly applicable within an economically oriented management of DQ, since both absolute and relative changes of the metric cannot be interpreted.

Table 1 demonstrates the weakness: To improve the value of correctness from 0.0 to 0.5, the corresponding distance function has to be decreased from ∞ to 1.0. In contrast, an improvement from 0.5 to 1.0 only needs a reduction from 1.0 to 0.0, i.e. only one failure has to be corrected. Summing up, it is not clear how an improvement of correctness (for example from 0.0 to 0.5) has to be interpreted.

Table 1. Improvement of correctness and necessary change of the distance function

Improvement of correctness	Necessary change of $d(w_I, w_R)$
0.0 → 0.5	∞ → 1.0
0.5 → 1.0	1.0 → 0.0

Besides correctness we take a closer look at the DQ dimension timeliness. Timeliness refers to whether the values of attributes are still up-to-date (for a literature review about existing definitions of timeliness see (Cappiello et al. 2004b)). The metric for timeliness shall deliver an indication (not a verified statement under certainty) whether an attribute value has changed in the real world since its acquisition and storage within the system or not. Therefore, Hinrichs proposed the following quotient (Hinrichs 2002):

$$Timeliness = \frac{1}{(mean\ attribute\ update\ time) \cdot (attribute\ age) + 1}$$

This quotient serves as a metric, which quantifies if the current attribute value is outdated. Related to the input factors taken into account, the quotient returns reasonable results: On the one hand, if the *mean attribute update time* is 0 (i.e. the attribute value never becomes out of date), timeliness is 1 (attribute value is up-to-date). If on the other hand *attribute age* is 0 (i.e. the attribute value is acquired at the instant of quantifying DQ) we get the same result. For higher values of *mean attribute update time* or *attribute age* the result of the metric approaches 0. I.e., that the (positive) indication (the attribute value is still corresponding to its real world counterpart) decreases.

However, this metric bears similar problems as the metric for correctness described above. Firstly, the value range [0; 1] is generally not covered, because we only get a value of 0 if the value of *mean attribute update time* or *attribute age* respectively is ∞ (R 1). In addition, by building a quotient the values are not interval scaled (R 2). Moreover, the metrics are hardly applicable within an economic oriented management of DQ, since both absolute and relative changes cannot be interpreted easily (R 3). I.e., the value of the metric cannot for instance be interpreted as a probability that the stored attribute value still corresponds to the current state in the real world. These facts hinder the evaluation of realized DQ measures ex post.

In contrast, another approach by (Ballou et al. 1998) defines the metric for timeliness as follows (the notation was slightly adapted):

$$Timeliness = \{\max[(1 - \frac{currency}{shelf\ \ life}), 0]\}^{s}$$

The *currency* of an attribute value is – in contrast to (Hinrichs 2002) – computed as follows: The time between the instant of quantifying timeliness and the instant of acquiring the attribute value is added to the age of the attribute value in the instant of acquiring it. This corresponds to the age of the attribute value at the instant of quantifying DQ. *Shelf life* is an indicator for the volatility of the attribute values. Thereby, a relatively high *shelf life* leads to a high result of the metric for timeliness and vice versa. By choosing *s* – which has to be assigned by experts – one can influence to which extent a change of the quotient (*currency*/*shelf life*) affects the result of the metric. Thereby it is possible to adapt the computation to the attribute considered and to the particular application (R 5).

It seems that it is the aim of (Ballou et al. 1998) to derive mathematical relations. They do not focus on getting values of the metric which are interpretable within an economic oriented management of DQ (cp. R 3) and easily understandable for instance by business departments (i.e. a marketing division in case of a CRM campaign). The results of their metric are only interpretable as the probability that the attribute value in the information system still corresponds to its real world counterpart in the case of $s = 1$ (in this case, a uniform distribution is assumed). For $s \neq 1$, the result of the metric cannot be regarded as a probability. I.e., that applying the exponent *s* worsens the interpretability of the results (cp. R 3) and they are no longer interval scaled (R 2).

Based on this literature review – that illustrates a gap in scientific literature –, we design metrics for the dimensions correctness and timeliness in the next section.

Design of Data Quality Metrics

Firstly, we consider the dimension correctness: Again, w_I is an attribute value and w_R the corresponding attribute value in the real world. $d(w_I, w_R)$ is a domain-specific distance function quantifying the difference between w_I and w_R. We want to assure the metric being normalized to the interval [0; 1] (R 1), without using a quotient. In contrast to the metric defined by Hinrichs, we therefore use a different functional equation and a distance function that is normalized to the interval [0; 1]. Examples for such distance functions are $d_1(w_I, w_R) := \begin{cases} 0 & \text{if } w_I = w_R \\ 1 & \text{else} \end{cases}$, which is independent of the field of application, $d_2(w_I, w_R) := \left(\frac{|w_I - w_R|}{\max\{|w_I|, |w_R|\}} \right)^{\alpha}$ with $\alpha \in \Re^+$ (normalized to the interval [0; 1]) for numeric, metrically scaled attributes and edit distance, Levenshtein distance or Hamming distance for strings (also in their well-known normalized type). The metric on the level of attribute values is therefore defined as follows:

(1) $$Q_{Corr.}(w_I, w_R) := 1 - d(w_I, w_R)$$

An example demonstrates how the metric works: Before starting a mailing campaign, the correctness of the attributes “postal code” and “house number” shall be evaluated. Based upon the distance function $d_2(w_I, w_I)$, the application of the metric can be illustrated as follows: First, to assure the adaptivity according to (R 5) on the level of attribute values the parameter α has to be chosen. In case of the attribute “postal code” even small deviations shall be penalized, because a deviation of only 1% (e.g. the postal codes 80000 (=w_I) and 79200 (=w_R)) hinders the delivery of a mailing. Therefore, the distance function has to react even on small deviations in a sensitive way. That is why α will be chosen such that α<1 (leading to an over-proportional increase of the function if $\frac{|w_I - w_R|}{\max\{|w_I|,|w_R|\}} << 1$). E.g. choosing α=0.01 – in case of a deviation of 1% – leads to a value of the metric for correctness of only 4.5%, which means that an attribute is classified as quite incorrect even in the case of small errors. In contrast, when considering the correctness of the attribute “house number”, smaller deviations are not so critical. This is due to the fact that the delivery of a mailing is still possible. In this case, the distance function ought to tolerate smaller deviations. Therefore, one will choose $\alpha > 1$ (under-proportional increase of the function if $\frac{|w_I - w_R|}{\max\{|w_I|,|w_R|\}} << 1$). For α=1.50 and a deviation of 1%, by using the metric we derive a value of 99.9%. After choosing α, both the value of the distance function and the DQ metric in formula (1) have to be computed.

In order to ensure the interpretability of the resulting values, we avoided a quotient and decided to use the functional term (1). Therefore, on the one hand the metric is interpretable referring to the related distance function. On the other hand the term (1) ensures that the value range of [0; 1] (following (R 1)) is met. The example introduced in the last section illustrates this. We use again the Levenshtein distance, which is normalized by means of dividing by the number of characters of the longer string:

$$Q_{Corr.}(\text{"Eissonhour"},\text{"Eisenhower"}) = 1 - d^{norm.}_{Lev.}(\text{"Eissonhour"},\text{"Eisenhower"}) = 1 - \frac{4}{10} = 60.0\%$$

$$Q_{Corr.}(\text{"Bird"},\text{"Hunt"}) = 1 - d^{norm.}_{Lev.}(\text{"Bird"},\text{"Hunt"}) = 1 - \frac{4}{4} = 0.0\%$$

Figure 3: Quantifying correctness by means of the designed metric

The example illustrates that the metric proposed – in contrast to Hinrichs’ approach – quantifies the attribute values’ correctness in the two cases quite differently: Considering “Eissonhour” and “Eisenhower”, the value of the metric for correctness is 60.0%, whereas looking at “Bird” and “Hunt” we derive a value of 0.0%, because both strings do not have a single character in common. Moreover, the values of the metric are (in case of an adequate normalized distance function) interval scaled (R 2). This ensures – in combination with meeting requirement (R 3) – that the metric is applicable within an economic management of DQ: For instance to increase the value of the metric by 0.5, the value of the distance function has to be decreased by 0.5, regardless whether the metric should be raised from 0.0 to 0.5 or from 0.5 to 1.0 (cp. Table 2).

Table 2. Improvement of correctness and necessary change of the distance function

Improvement of correctness	Necessary change of $d(w_I, w_R)$
0.0 → 0.5	1.0 → 0.5
0.5 → 1.0	0.5 → 0.0

To meet the requirement of aggregation (R 4) the metric is constructed “bottom up”. I. e., a metric on level *n+1* (e.g. correctness on the level of tupels) is based on the corresponding metric on level *n* (e.g. correctness on the level of attribute values). This also ensures that the designed metric allows quantifying DQ on the levels of attribute values, tupels, relations and database. Hence, the DQ metric on the level of tupels is now designed based upon the metric on the level of attribute values. Assume *t* to be a tupel with attribute values $t.A_1, t.A_2,\ldots, t.A_n$ for the attributes $A_1, A_2,\ldots, A_n$ and $e.A_1, e.A_2,\ldots, e.A_n$ being the corresponding attribute values *e* of the real world entity. Due to requirement (R 5), the relative importance of the attribute A_i with regard to correctness can be weighted with $g_i \in [0; 1]$. Consequently, the metric for correctness on the level of tupels – based upon (1) – can be written as:

(2) $$Q_{Corr.}(t,e) := \frac{\sum_{i=1}^{n} Q_{Corr.}(t.A_i, e.A_i)g_i}{\sum_{i=1}^{n} g_i}$$

The correctness of a tupel t is based on the correctness of the attributes involved. On the level of relations the correctness of a relation R (or a relational view) can be defined via the arithmetical mean of the values of the metric for the tupel $t_j \in R$ (j=1, 2, …, $|R|$) as follows (if R is a non-empty relation and E the corresponding set of entities in the real world):

(3) $$Q_{Corr.}(R,E) := \frac{\sum_{j=1}^{|R|} Q_{Corr.}(t_j, e_j)}{|R|}$$

Assume D being a database that can be represented as a disjoint decomposition of the relations R_k (k=1, 2, …, $|R|$). I. e., the whole database can be decomposed into pairwise non-overlapping relations R_k, so that each attribute of the database is assigned to exactly one of the relations. Formally noted: $D=R_1 \cup R_2 \cup \ldots \cup R_{|R|}$ and $R_i \cap R_j = \emptyset \ \forall i \neq j$. Moreover, R is the corresponding modeled part of the real world, where E_k represents the set of entities associated with R_k. Then the correctness of a database D can be (based on the correctness of the relations R_k (k=1, 2, …, $|R|$)) defined as:

(4) $$Q_{Corr.}(D,R) := \frac{\sum_{k=1}^{|R|} Q_{Corr.}(R_k, E_k)g_k}{\sum_{k=1}^{|R|} g_k}$$

Whereas (Hinrichs 2002) defines the correctness of a database by means of an unweighted arithmetical mean, the weights $g_k \in [0; 1]$ allow to incorporate the relative importance of each relation depending on the given context (R 5). According to the approach of Hinrichs, relations that are not important for the given scope are equally weighted to relations of high importance. In addition, the resulting measurement depends on the disjoint decomposition of the database into relations. This makes it difficult to evaluate the correctness of a database objectively. E. g., a relation R_k with $k \neq 2$ is weighted relatively with $1/n$ when using the disjoint decomposition $\{R_1, R_2, R_3, \ldots, R_n\}$, whereas the same relation is only weighted with $1/(n+1)$ when using the disjoint decomposition $\{R_1, R_2', R_2'', R_3, \ldots, R_n\}$ with $R_2' \cup R_2'' = R_2$ and $R_2' \cap R_2'' = \emptyset$.

The designed metric enables us to quantify correctness at each level. Comparing the attribute values stored in the information system to their real world counterparts is crucial to do so. The metric can be computed automatically, if the real world counterpart of each considered attribute value is known, which is achieved by conducting a survey. This is hard to do at reasonable costs for huge sets of data. That is why we propose to consider only a sample of attribute values that is representative for the whole dataset and survey the real world counterpart of these attribute values. Thereby, conclusions can be drawn from the sample for the whole dataset. Thus, we get an estimator for $Q_{Corr.}$. E.g., it is possible to buy up-to-date address data from external sources for a sample of the customer base. These bought addresses can be compared to the address data stored in the information system. The value of the metric Q_{Corr} can then be used as an estimator for the value of the correctness of the whole set of address data. This estimation can be a starting point for taking DQ measures in the next step.

After discussing correctness, we focus on timeliness in the following and, in a first step, design a metric on the level of attribute values. The results on this level can be aggregated on the levels of tupels, relations and database by analogy with the metric for correctness.

During the process of designing the metrics, we analyzed common approaches (e.g. a parameterized root function) for normalizing the range of the values (in order to meet (R 1)). They assured the values being in the interval [0; 1], but the other requirements were still not met. Aiming especially at the results being interval scaled (R 2) and interpretable (R 3) and in order to enable an automatic measurement of timeliness (to a large extent – see table 3) (R 6), we suggest an approach which is founded on probability theory. Thereby, a difference between two values of the metric is meaningful, since the results of the metric can be interpreted as a probability of an attribute value still corresponding to its real world counterpart.

In the following we assume the underlying attribute values‘ shelf-life to be exponentially distributed. The exponential distribution is a typical distribution for lifetime, which has proven its usefulness in quality management (especially for address data etc.). The density function *f*(*t*) of an exponentially distributed random variable is noted – depending on the decline rate *decline*(*A*) of the attribute *A* – as follows:

$$f(t)=\begin{cases} decline(A)\cdot \exp(-decline(A)\cdot t) & \text{if } t\geq 0 \\ 0 & \text{else} \end{cases}$$

Using this density function one can determine the probability of an attribute value losing its validity between t_1 and t_2. The surface limited by the density function within the interval [$t_1; t_2$] represents this probability. The parameter *decline*(*A*) is the decline rate indicating how many values of the attribute considered become out of date on average within one period of time. E. g., a value of *decline*(*A*)=0.2 has to be interpreted as follows: on average 20% of the attribute *A*'s values lose their validity within one period of time. Based on that, the distribution function *F*(*T*) of an exponentially distributed random variable indicates the probability of the attribute value considered being outdated at *T*. It is denoted as:

$$F(T)=\int_{-\infty}^{T} f(t)dt=\begin{cases} 1-\exp(-decline(A)\cdot t) & \text{if } T\geq 0 \\ 0 & \text{else} \end{cases}$$

Based on the distribution function *F*(*T*), the probability of the attribute value being valid at *T* can be determined in the following way:

$$1-F(T)=1-(1-\exp(-decline(A)\cdot T))=\exp(-decline(A)\cdot T)$$

We use this equation to define the metric on the level of attribute values. Thereby *age*(*w*, *A*) denotes the age of the attribute value *w*, which is computed by means of two factors: the instant when DQ is quantified and the instant of data acquisition. In contrast to (Ballou et al. 1998) we can use the instant of data acquisition because the exponential distribution is memoryless. This also enables an automated measurement (R 6). Moreover, the decline rate *decline*(*A*) of attribute *A*'s values can be determined statistically (see next section for examples). The metric on the level of an attribute value is therefore defined as:

(5) $$Q_{Time.}(w,A):=\exp(-decline(A)\cdot age(w,A))$$

$Q_{Time.}(w, A)$ denotes the probability that the attribute value is still valid. This interpretability (R 3) is an advantage compared to existing approaches. By representing a probability, the metric (5) is normalized (R 1) and interval scaled (R 2).

For attributes that never change (as e. g. "date of birth"), we choose *decline*(*A*)=0, resulting in a value of the metric equal to 1: $Q_{Time.}(w,A)=\exp(-0\cdot age(w,A))=1$. Moreover, the metric is equal to 1, if an attribute value is acquired at the instant of quantifying DQ – i. e. *age*(*w*, *A*)=0: $Q_{Time.}(w,A)=\exp(-decline(A)\cdot 0)=1$. The re-collection of an attribute value is also considered as an update of an existing attribute value. In order to meet requirement (R 4) the metric for timeliness was designed in such way that its definition – similarly to the metric for correctness – on the level of tupels, relations and database bases on the metric on the next lower level.

Table 3 sums up the results and contains the steps necessary for quantifying timeliness and denotes (1) whether the measurement is objective or subjective, (2) whether the steps have to be done manually or can be automated and (3) how often they are accomplished when quantifying the timeliness of attribute values.

Table 3. Steps for quantifying timeliness

Step	objectivity/human involvement/frequency
1. Selection of data attributes to be evaluated (e. g. customer attributes within a marketing campaign)	subjective/manual/only once for all values of one attribute
2. Configuration of the metric:	
• Determination of the weights: The weights of the selected attributes – representing their importance for the context – have to be fixed (R 5)	subjective/manual/only once for all values of one attribute
• Determination of the distribution parameter: External or internal data has to be analyzed for estimating the decline rate of each selected attribute (R 3)	subjective or objective[2]/manual/only once for all values of one attribute
3. Measurement of timeliness for each attribute value (R 6). This can be performed by means of SQL DML-statements:	
• Computation of the age of the attribute values by means of meta data (instant of data acquisition) and the instant of evaluating DQ	objective/automated/for each attribute value
• Computation of the value of the metric	objective/automated/for each attribute value
• Aggregation of the values of the metric for the selected attribute values to the level of tupels (according to the weights) (R 4)	objective/automated/for each attribute value
• Processing of the metric results (e. g. to classify the customers according to the timeliness of their attributes)	objective/automated/for each attribute value

Case study: Application of the Metric for Timeliness

In this section the evaluation of the developed metric for timeliness is illustrated by a case study, i.e. we analyze whether the metric is applicable and meets the requirements adaptivity (R 5) and feasibility (R 6). Therefore, we applied the metric in a business environment within the campaign management of a major German mobile services provider. For reasons of confidentiality, the figures and data had to be changed and made anonymous. Nevertheless, the procedure and the basic results remain the same.

In the past, existing DQ problems often prohibited a correct and individualized customer addressing in mailing campaigns and led to lower campaign success rates. This problem occurred especially with prepaid contracts, since they do not guarantee customer contact at regular intervals (e.g. sending bills). Hence, the mobiles services provider cannot easily verify, whether these customers' contact data are still up-to-date. In the following, we consider a Prepaid2Postpaid campaign. Its aim was to submit an offer to the prepaid customers in order to make them switch to postpaid tariffs. In the considered campaign 143,000 customers with the prepaid tariff "*Mobile1000*" are offered to switch to the postpaid tariff "*Mobile2500*". "*Mobile2500*" is more profitable for the mobile services provider, since its contract period is fixed and it guarantees minimum sales. Due to the large number of customers that shall be addressed, using the metric for correctness is very time-consuming, since the address of each single customer must be verified before mailing the offer. In contrast, the metric for timeliness provides the advantage of giving a probability that indicates whether the address of a customer is still up-to-date without knowing under certainty whether it is indeed still correct. Its advantage is having the mostly automatable computation of the probability and thus being more cost-efficient. Hence, we focus on the metric for timeliness that was applied in the campaign as follows (due to space restrictions we can not illustrate the metric for correctness as well):

Firstly, the relevant attributes and their relative importance within the campaign had to be determined (R 5). The attributes "surname", "first name" and "address" ("street", "house number", "postal code" and "city" in detail) were considered important for delivering the offer to the customer by mail. Moreover, the customer's current tariff was essential, since it was the selection criterion within the campaign. Next, the relative weights of these four attributes

[2] The distribution parameter may be determined in an objective way, when their estimation is based on samples or statistical distributions (e.g. provided by Federal Statistical Offices) of historical data. In contrast, if such data are not available, we have to rely on (subjective) estimations by experts.

had to be specified according to their importance within the campaign. Since only those customers with the tariff "*Mobile1000*" should be addressed, the attribute "current tariff" was stated as most important. Therefore, it got the relative weight 1.0, which served as a reference base for the weights of the other attributes. The attribute "address" was considered next important, since without the address, the offer cannot be delivered to the customer. Nevertheless "address" was not given a weight of 1.0, but only 0.9, since parts of the address – as e. g. an up-to-date house number – are not indispensable for the offer's delivery. Accordingly, the attribute "surname" was weighted 0.9, since this attribute is also very important for the delivery, but if the surname changes (e.g. after a marriage), the old surname might – in some cases – still be known to the postal service. In contrast, the first name was considered less important. A wrong first name might annoy the customer, but it does generally not prevent the delivery. The attribute "first name" was nevertheless assigned the weight 0.2, since the mobile services provider did not want to affect existing customer relationships. In the project, the relative weights were set in collaboration with the marketing department. Alternatively, the impact of the particular attributes on the goal can be analyzed by means of samples (e.g. how important is an up-to-date surname compared to an up-to-date address with respect to the success of the campaign?), for estimating the weights more objectively.

Considering this procedure, we have to analyze how the values of the metric react to changes of the weights (sensitivity and robustness of the results). This is ambivalent: On the one hand, the values of the metric should and must change due to a significant change of the weights. On the other hand, smaller changes of the weights (e.g. because of estimation errors) should not lead to very different values of the metric. According to a sensitivity analysis, the metric reacts robustly to smaller changes of an individual weight g_i. Based on the example in Table 3, we found that an isolated variation of an individual weight for instance by +/- 10% causes a change of the result of the metric by less than 0.5%. This was supported by an analysis of the data of the German mobile services provider: An isolated variation of an individual weight by +/-10% changed the assignment to the intervals [0; 0.1],]0.1; 0.2], …,]0.9; 1] (see next paragraphs) for less than 1% of the customers. In contrast, if a mutual variation is intended, the result of the metric reacts significantly.

In the next step, the decline rate *decline*(A_i) had to be specified for each attribute, which can be done for instance by means of statistics. Regarding the attributes "surname" and "address", empirical data from the Federal Statistical Office of Germany considering marriages/divorces and the frequency of relocation were taken into account. Thereby decline rates of 0.02 for the attribute "surname" (i. e. on average, 2% of all customers change their surname p.a.) and 0.1 for the attribute „address" could be determined. If no such third party data is available, the decline rate of the attribute "address" can be estimated by means of own (historical) data or samples. E.g., in the case of frequency of relocation, it would be possible to draw a sample of the customer base. After having surveyed the average duration of validity of the customer addresses in the sample (i.e., how long does a customer live in the same habitation on average?), the parameter *decline(A)* of the metric for timeliness can be set by means of the unbiased estimator $\left(\frac{1}{\text{average duration of validity of the addresses}}\right)$ for the probability distribution. The decline rate of the attribute "first name" was assumed as 0.0 since the first name usually remains the same. In contrast the decline rate of "current tariff" was estimated based on historical data from the mobile services provider as 0.4. Last, the age of each attribute had to be computed automatically (R 6) by means of two parameters: the instant when DQ is quantified and the instant of data acquisition, which is stored as meta data in the data records. It allows to determine the variable *age*($T.A_i$, A_i). Table 4 sums up the relevant values.

Table 4. Quantifying timeliness by means of the metric (example for a customer tupel)

A_i	surname	first name	address	current tariff
g_i	0.9	0.2	0.9	1.0
age($T.A_i$, A_i) [year]	0.5	0.5	2	0.5
decline(A_i) [1/year]	0.02	0.00	0.10	0.4
$Q_{Time.}(T.A_i, A_i)$	**0.99**	**1.00**	**0.82**	**0.82**

The value of the metric on the level of tupels is computed via aggregation of the results on the level of attribute values, considering the weights g_i:

$$Q_{Time.}(T, A_1, ..., A_4) = \frac{0.99 \cdot 0.9 + 1 \cdot 0.2 + 0.82 \cdot 0.9 + 0.82 \cdot 1}{0.9 + 0.2 + 0.9 + 1} \approx 0.882$$

Hence, the resulting value of the metric for timeliness is 88.2% for the exemplary tupel. This means that the tupel for the given application (promoting a tariff option) is up-to-date at a level of 88.2%. Before applying the metrics to the current campaign, a very similar campaign was analyzed which was performed three months earlier.

The former campaign addressed 82,000 customers who had been offered a tariff switching, too. The success rate was about 8.5% on average, i. e. about 7,000 customers could be convinced to switch their tariff. The metric for timeliness $Q_{Time.}(T.A_1, \ldots, A_4)$ was computed for all customers addressed in the former campaign. Afterwards, the customers were classified according to the value of the metric, i. e. each customer was assigned to one of the intervals [0; 0.1],]0.1; 0.2], …,]0.9; 1]. For each interval the percentage of customers that accepted the offer (campaign success rates) was determined. Figure 4 depicts the results.

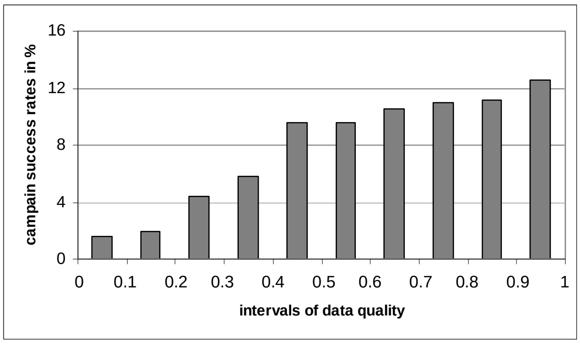

Figure 4: Success rate of a former campaign depending on the metric for timeliness

The figure illustrates: The more timely the attributes of a customer, the higher the success rate of the campaign (note that this does not mean that the success rate depends only on the timeliness of the attributes mentioned above). E.g., the success rate within the interval]0.2; 0.3] is only 4.4%, whereas in the interval]0.9; 1] it is 12.6%. This is not surprising, since customers with an outdated address for instance do not even have the possibility to accept the offer – because they quite simple did not receive it.

These results become still more interesting when looking at the current campaign. Its target group consists of 143,000 customers in total (all customers with the tariff *"Mobile1000"*). Assuming the success rates shown in Figure 4 (because the former and the new campaign were very similar), it does not make sense – from an economic point of view – to address customers with a value of the metric for timeliness below 0.3. E.g. in the new campaign the number of customers within the interval]0.2; 0.3] is 12,500. Taking into account the expected success rate of only 4.4% (about 550 customers) for these customers, the costs of mailing are higher than the (expected) additional revenues resulting from tariff switching. Only starting from the interval]0.4; 0.5] the additional revenues outweigh the mailing costs (see profit (without buying addresses) on the right side of the chart in Figure 7). The profitability of the campaign can be increased by 30 percentage points by addressing only those customers with a value of the metric for timeliness higher than 0.4. This is due to the fact that mailing costs are reduced and the (expected) average success rates are improved.

However, the economic management of DQ presented in this paper does not stop at this point. Indeed, the profitability of the campaign was improved. But it remains dissatisfying that customers who might switch their tariff cannot accept the offer, because it cannot be delivered to them (due to an outdated address). It was analyzed, if for instance buying external data – as one possible DQ measure besides others – could help to solve this problem. E.g. firms like German Postal Service offer up-to-date address data they acquire from forwarding requests of relocators. The question was now, whether this measure should be taken or not, since buying address data raises costs on the one hand, on the other hand it improves DQ and therefore the campaign's success rate. The procedure for making up a decision was as follows: Firstly, for each interval the costs for buying addresses of customers were calculated. This could be done easily, since the number of customers per interval was known and firms (as German Postal Service) charge a fixed price for each updated address. These costs have to be compared to the revenues resulting from a higher success rate (the offer can now be delivered to the customer). Calculating the higher success rates, the degree of improving DQ (by buying addresses) has to be determined. The latter can be done via the formula $Q_{Time.}(T.A_i, A_i)$. By means of the metric the (improved) success rate can be estimated by assuming the success rates of the former, similar campaign (see chart on the left side in Figure 5). The improved success rates can be used for calculating the additional revenues. On this basis we calculated the additional profit by buying addresses. Figure 5 illustrates this (right chart).

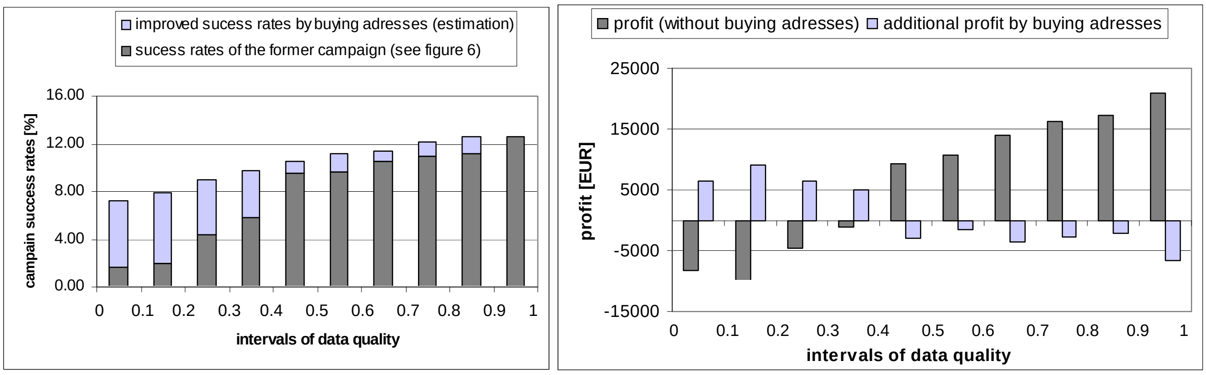

Figure 5: Campaign success rates and profits depending on the metric for timeliness

The results are twofold: One the one hand, performing the campaign without buying addresses – from an economic point of view – only makes sense within the interval]0.4; 1]. Within this interval buying addresses is senseless, because the costs of buying addresses are higher than the expected revenues (resulting in a negative additional profit). On the other hand, performing the campaign does not make sense for customers from the interval [0; 0.4] at all, if we do not take into account the possibility of buying addresses. However, buying addresses results in additional profits for this interval. Indeed, the entire profit (total expected revenues minus costs of mailing and buying addresses) is only positive within the interval]0.2; 0.4] (considering the interval [0; 0.4]). Therefore, buying addresses solely makes sense for about 25,400 customers within this interval. The mobile services provider tried to buy addresses for these 25,400 customers from an external firm and got address data for approx. 20,000 customers. These addresses were compared to the ones stored in the database. The analysis revealed that only about 3,470 addresses of customers assigned to the interval]0.2; 0.3] (12,500 customers in total) and approx. 4,250 addresses of customers assigned to the interval]0.3; 0.4] (12,900 customers in total) were up-to-date. This illustrates that timeliness (in terms of a probability) estimated ex ante was accurate on average.

In the next step, the offer should be sent to those customers in the interval]0.2; 1]. However, as a precaution, the mobiles services provider decided to address all 143,000 customers including those with a low value of timeliness of their addresses (interval [0; 0.2]). This turned out to be unreasonable from an economic point of view, because the estimated success rates were – as already mentioned – a good base for the ex post success rates (the maximal difference was ± 0.6%). E.g., the ex post success rate of the customers in the interval]0.1; 0.2] was indeed only at 2.0% instead of the estimated 1.9%. So, the campaign was not profitable within the interval [0; 0.2].

Besides this example for the metric's application resulting in both lower campaign and measures costs, several DQ analyses were conducted for raising profits. By applying the metrics, the mobile services provider was able to establish a connection between the results of quantifying DQ and the success rates of campaigns. Thereby the process for selecting customers was improved significantly for the campaigns of the mobile services provider, since campaign costs could be cut down, too. Moreover, the mobile services provider can take DQ measures more efficiently and estimate the economic benefit more accurately.

Summary

In this paper it was analyzed how DQ dimensions can be quantified in a goal-oriented and economic manner. The aim was to design innovative metrics for the DQ dimensions correctness and timeliness. In cooperation with a major German mobile services provider, the metrics were applied in the case of campaign management and they proved appropriate. We choose this case study, since it allows to illustrate and to measure the impact of poor and improved DQ, respectively. The resulting improvements can be isolated and economically interpreted by comparing the estimated (ex ante) success rate of campaigns with the realized one (ex post). Moreover, the realization (ex post) depends on the behavior of an “(independent) third party" outside the firm (the customers). Therefore, it is objectively verifiable to a large extent. However, we can apply the metrics to other cases, even outside the field of CRM: They can be applied in domains, in which data values get outdated over time. They can also be used, whenever data values can become incorrect during their acquisition or changing. Applying the metrics is reasonable, if a dataset serves as a base for decisions or is processed, provided that the assumptions above hold. From this point of view, they can also be used in management, production or logistic processes (cp. examples given in Ballou et al. 2002). In contrast to existing approaches, the metrics were designed according to important requirements like interpretability and feasibility. They allow quantifying DQ and thereby represent the foundation for economic analyses. Furthermore, the proposed metric for timeliness enables an objective and automated measurement during important steps of the quantifying process. Thereby, the results fill a gap in both, science and practice.

References

Ballou, D. P., Wang, R. Y., Pazer, H., and Tayi, G. K. “Modeling information manufacturing systems to determine information product quality,” *Management Science* (44:4), 1998, pp. 462-484.

Campanella, J. *Principles of quality cost*, ASQ Quality Press, Milwaukee, 1999.

Cappiello, C., Francalanci, Ch., and Pernici, B. “Data quality assessment from the user’s perspective,” in *Proceedings of the 2004 international workshop on Information quality in information systems,* Paris, 2004a, pp. 68-73.

Cappiello, C., Francalanci, Ch., and Pernici, B. “Time-Related Factors of Data Quality in Multichannel Information Systems,” *Journal of Management Information Systems* (20:3), 2004b, pp. 71-91.

Cappiello, C., Francalanci, Ch., Pernici, B., Plebani, P., and Scannapieco, M. “Data Quality Assurance in Cooperative Information Systems: A multi-dimensional Quality Certificate,” in *International Workshop on Data Quality in Cooperative Information Systems,* T. Catarci (ed.), Siena, 2003, pp. 64-70.

English, L. *Improving Data Warehouse and Business Information Quality*, Wiley, New York, 1999.

Eppler, M. J. *Managing Information Quality*, Springer, Berlin, 2003.

Even, A., and Shankaranarayanan, G. “Utility-Driven Assessment of Data Quality,” *The DATA BASE for Advances in Information Systems,* (38:2), 2007, pp. 75-93.

Even, A., and Shankaranarayanan, G. “Value-Driven Data Quality Assessment,” in *Proceedings of the 10th International Conference on Information Quality,* Cambridge, 2005.

Feigenbaum, A. V. *Total quality control*, McGraw-Hill Professional, New York, 1991.

Heinrich, B., and Helfert, H. “Analyzing Data Quality Investments in CRM – a model based approach,” in *Proceedings of the 8th International Conference on Information Quality,* Cambridge, 2003.

Hevner, A. R., March, S. T., Park, J., and Ram, S. “Design Science in Information Systems Research,” *MIS Quarterly* (28:1), 2004, pp. 75-105.

Hinrichs, H. *Datenqualitätsmanagement in Data Warehouse-Systemen,* doctoral thesis, Oldenburg, 2002.

Jarke, M., and Vassiliou, Y. “Foundations of Data Warehouse Quality – A Review of the DWQ Project,” in *Proceedings of the 2nd International Conference on Information Quality*, Cambridge, 1997.

Juran, J. M. “How to think about Quality,” in *Juran’s Quality Handbook*, McGraw-Hill, New York, 2000.

Lee, Y. W., Strong, D. M., Kahn, B. K., and Wang, R. Y. “AIMQ: a methodology for information quality assessment,” *Information & Management* (40), 2002, pp. 133-146.

Meta Group *Data Warehouse Scorecard*, Meta Group, 1999.

Pipino, L. L., Lee, Y. W., and Wang, R. Y.: “Data Quality Assessment,” *Communications of the ACM* (45:4), 2002, pp. 211-218

Redman, T. C. *Data Quality for the Information Age*, Arctech House, Norwood, 1996.

Redman, T. C. “The Impact of Poor Data Quality on the Typical Enterprise,” *Communications of the ACM* (41:2), 1998, pp. 79-82.

SAS Institute *European firms suffer from loss of profitability and low customer satisfaction caused by poor data quality*, Survey of the SAS Institute, 2003.

Shank, J. M., and Govindarajan, V. “Measuring the cost of quality: A strategic cost management perspective,” *Journal of Cost Management* (2:8), 1994, pp. 5-17.

Strong, D. M., Lee, Y. W., and Wang R. Y. “Data quality in context,” *Communications of the ACM* (40:5), 1997, pp. 103-110.

Teboul, J. *Managing Quality Dynamics*, Prentice Hall, New York, 1991.

Wang, R. Y., Storey, V. C., and Firth, C. P. "A Framework for analysis of data quality research," *IEEE Transaction on Knowledge and Data Engineering* (7:4), 1995, pp. 623-640.