Assessing data currency – a probabilistic approach

Bernd Heinrich and Mathias Klier[1]

Department of Information Systems, Production and Logistics Management, University of Innsbruck, Austria

Abstract.

The growing relevance of data quality has revealed the need for adequate measurement. As time aspects are extremely important in data quality management, we propose a novel approach to assess data currency. Our metric, which is founded on probability theory, enables an objective and widely automated assessment for data liable to temporal decline. Its values are easy to interpret by business users. Moreover, the metric makes it possible to analyse the economic impacts of data quality measures like data cleansing and can therefore build a basis for an economic management of data quality. The approach can be applied in various fields of application where the currency of data is important. To illustrate the practical benefit and the applicability of the novel metric, we provide an extensive realworld example. In cooperation with a major German mobile services provider, the approach was successfully applied in campaign management and led to an improved decision support.

Keywords: data quality; data quality assessment; data quality metrics

1. Introduction

Both the benefit and the acceptance of information systems depend heavily on the quality of data provided by these systems [1-3]. Executives and employees need high-quality data in order to perform business, innovation, and decision-making processes properly [4, 5]. It is, therefore, not surprising that bad data quality may lead to wrong decisions and correspondingly high costs. According to an international survey, 75 percent of all respondents have already made wrong decisions due to incorrect or outdated data. In addition, they and their staff spend up to 30 percent % of their working time on checking the quality of data provided [6]. Here, ensuring completeness, correctness and currency of data – such properties are known as data quality dimensions [7] – still remains an important problem for many companies and public institutions [8-13]. But how good is an organization's data quality? To answer this important question, well-founded and applicable metrics are needed [8, 14, 15]. In addition, assessing data quality (e.g. master data) is essential for analysing the economic effects of bad or improved data quality as well as for planning data quality measures in an economic manner. In this paper, we present a novel metric for currency, as empirical investigations reveal that time aspects are extremely

[1] *Correspondence to: Bernd Heinrich, University of Innsbruck, Department of Information Systems, Production and Logistics Management, Universitaetsstrasse 15, A-6020 Innsbruck, Austria, Email: bernd.heinrich@uibk.ac.at*

important in data quality management (cf. [16, 17]). This probabilistic approach enables an objective and widely automated data quality assessment and its values are easy to interpret by business users. Moreover, the metric can be applied for data liable to temporal decline. Therefore, the possible fields of application are manifold. They range from data in production and logistic processes (e.g. data exchange in value networks) to data in public institutions (e.g. user account data in libraries). To illustrate this novel approach, we single out one application and provide a real world example from the customer relationship management (CRM) context. In cooperation with a major mobile services provider this approach was applied in campaign management and improved both success rates and profits. In this context, it is quite easy to illustrate the practical benefit of the novel metric. However, the approach is not limited to this domain.

The remainder of the paper is organized as follows. The next section briefly describes the problem context and the state of the art. In Section 3, six requirements for data quality metrics for currency are derived from the literature. These requirements serve as guidance for the design of the novel metric to assess the currency of data, which is founded on probability theory, in Section 4. Section 5 illustrates the applicability of the novel metric by means of an extensive real-world example. In Section 6, the findings of the application and an ex-post analysis demonstrate the practical benefit of the novel approach and serve as an evaluation. The last section summarizes our findings and critically reflects on the results.

2. Problem context and state of the art

In literature there are two different perspectives of quality: quality of design and quality of conformance [18-20]. The former denotes the degree of correspondence between the users' demand and the specification of the information system (which is, for example, specified by means of data schemata). In contrast, quality of conformance represents the degree of correspondence between the data values stored in a database and the corresponding real world counterparts (for example, are the stored data values still up-to-date?). Figure 1 illustrates these two perspectives.

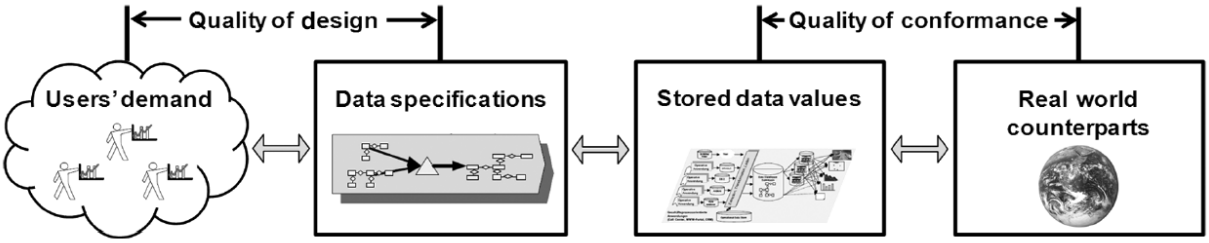

Fig. 1. Quality of design vs. quality of conformance.

The distinction between quality of design and quality of conformance is also important in the context of assessing data quality: It separates the (mostly subjective) analysis of the correspondence between the users' demand (e.g., derived by means of an information demand analysis) and the specified data schemata from the assessment of the correspondence between the stored data values and the corresponding real-world counterparts, which is more objective. In the following we focus on Quality of Conformance which is mainly related to data values and more independent of particular users' demand in specific business situations. According to Redman, the data quality dimensions correctness, completeness, consistency and currency are most important in this context [21]. These data quality dimensions have been discussed from both a scientific and a practical point of view in many publications [9, 22-25]. Empirical investigations reveal that time aspects are extremely important in data quality management [16, 17], while the data quality dimension currency is extensively discussed in the

scientific literature [5, 8, 14, 26, 27]. Hence, we focus on this data quality dimension in the following and seek to assess the quality of datasets by means of a metric for currency.

In literature no standard definition has been established for the data quality dimension currency [26]. Often this dimension is also referred to as timeliness [27] and sometimes it is even seen as a part of timeliness [8]. Table 1 provides selected definitions.

Table 1. Definitions with respect to the data quality dimension currency

Author(s)	Term and definition
Ballou and Pazer (1985)	Timeliness: 'The recorded value is not out of date' [28].
Ballou, Wang, Pazer, and Tayi (1998)	Timeliness: 'The timeliness of a raw or primitive data unit is governed by two factors. The first, currency, refers to the age of the primitive data units used to produce the information products. The second, volatility, refers to how long the item remains valid' [8].
Batini and Scannapieco (2006)	Timeliness: 'Timeliness expresses how current data are for the task at hand' [22].
Cappiello, Francalanci, and Pernici (2003)	Currency: 'Currency is defined as the degree to which data are up to date in a given operational database' [26].
Nelson, Todd, and Wixom (2005)	Currency: 'The degree to which information is up to date, or the degree to which the information precisely reflects the current state of the world that it represents' [27].
Redman (1996)	Currency: 'Currency refers to a degree to which a datum in question is up-to-date. A datum value is up-to-date if it is correct in spite of possible discrepancies caused by time-related changes to the correct value' [21].

Following the main consensus of these definitions, currency expresses whether an attribute value is still up-to-date. This means that an attribute value, which was correct when it was stored in a database, still corresponds to the current value of its real world counterpart at the instant when data quality is assessed. In other words, the attribute value has not become outdated (due to temporal decline). In contrast to correctness, assessing currency does not necessarily require a real world test. This is a huge advantage of the dimension currency, since comparing attribute values to their real world counterparts – which is necessary for assessing correctness – is often far too time-consuming and cost-intensive. In the case of large datasets such a real world test is not at all practicable (for example, 189,000 stored customer datasets considered within the campaign of the mobile services provider). Instead, a metric for currency shall provide an indication, not a verified statement under certainty, whether an attribute value has changed in the real world since its acquisition and storage in the database.

In recent years, researchers have taken an economic viewpoint for data quality management and developed methodologies to assess data quality accordingly [5, 29, 30]. Hence, the literature already provides approaches for assessing data quality in general and the data quality dimension currency in particular. In the following paragraphs, we give a brief overview of selected previous work in that field.

A well-known approach for assessing data quality is the AIM Quality (AIMQ) methodology [25]. It encompasses three elements: (1) a product service performance model which consolidates the most important data quality dimensions into four quadrants; (2) a questionnaire to assess data quality; and (3) analysis techniques for interpreting the values resulting from the data quality assessment. Even if the AIMQ methodology is a very important contribution in the field of data quality assessment, the authors do not seek to develop formally noted metrics (for currency). However, they use a questionnaire to assess data quality and to determine the best areas for data quality improvement. Besides this major scientific contribution, we briefly want to mention the widely spread works by English and Redman. English proposes the Total Quality data Management (TQdM) methodology which has been successfully applied in a large number of cases [9]. The methodology, sometimes also referred to as Total Information Quality Management (TIQM), follows the concepts of Total Quality Management and

comprises techniques for assessing data quality. Redman presents a process-oriented approach based on the concept of statistical quality control [21]. Both authors focus on practical aspects and provide interesting managerial principles. However, neither seeks to develop formally noted metrics to assess the data quality dimension currency. In contrast, Pipino et al. and Heinrich et al. focus on assessing data quality by means of metrics. Pipino et al. describe general principles that can help organizations develop usable data quality metrics [15]. In this context, the authors provide important insights in using the three major functional forms of simple ratio, min or max operators, and weighted average when developing (formally noted) data quality metrics. We will refer to this work when developing our approach in Section 4. Heinrich et al. in particular address the data quality dimension currency [14]. However, the authors do not seek to develop a concrete formally noted metric to assess the currency of data, but rather propose a general procedure to develop metrics.

Beyond those approaches discussed above, there are concrete metrics which are designed to assess the currency of data, are formally noted, and are based for the most part on a quality of conformance definition (Figure 1). In the following, we analyse these metrics.

One of the first and most renowned contributions in this field is the approach by Ballou et al. [8]. Whereas the authors propose a metric for timeliness, this approach is very important in our context as Ballou et al.'s understanding of timeliness is very similar to our definition of the data quality dimension currency (Table 1). The metric by Ballou et al. is shown below. For reasons of consistency, the notation has been slightly adapted:

$$Currency := \left\{ \max\left[1 - \frac{age\ of\ \ attribute\ value}{shelf\ \ life}; 0 \right] \right\}^{s} \tag{1}$$

The parameter *age of attribute value* is defined as follows: the time period between the assessment of currency and the acquisition of the attribute value (i.e. time of the data in the information system) is added to the age of the attribute value at the instant of acquiring it. This corresponds to the age of the attribute value at the instant of assessing currency. The parameter *shelf life* represents the volatility and is defined as the maximum length of time during which the attribute value in question remains valid. Thus, a high value for *shelf life* related to *age of attribute value* results in a high value of the metric for currency and vice versa. The exponent $s > 0$ allows control of the sensitivity of the values of the metric to the ratio *age of attribute value*/*shelf life*. This way, the metric can be adapted to the context of a particular application. In addition, for comparison purposes it is a major advantage of the metric that Ballou et al. assess currency on a continuous scale from 0 to 1. Thereby, a value of 1 represents perfectly good, and a value of 0 perfectly bad, data quality regarding the dimension currency, respectively.

Hinrichs defines the following metric to assess the currency of data, which shall provide an indication whether an attribute value is still up-to-date [31]:

$$Currency := \frac{1}{\left(mean\ attribute\ update\ \ frequency \cdot age\ of\ \ attribute\ value\right) + 1} \tag{2}$$

The parameter *mean attribute update frequency* denotes how often attribute values are updated on average within a certain period of time (e.g. three times per year). In contrast to the metric defined by Ballou et al., the parameter *age of attribute value* is defined as the time period between the assessment of currency and the acquisition or update of the attribute value (i.e. the period that the data are stored in the information system). Hence, the age of the attribute value at the instant of acquiring it (i.e. the period between the instant when the real world event occurred and the instant when the data were entered in the information system) is not necessary to calculate the value of the metric. That can be an advantage of this approach since the parameter *age of attribute value* as defined by Hinrichs is available in most cases and can often be determined automatically from the metadata. In addition, the values of the metric are normalized to the interval [0; 1] as suggested by Ballou et al. [8].

Even and Shankaranarayanan propose a utility-based approach for assessing currency [5]. The metric represents a function of the parameter *age of attribute value* (defined as the time period between the assessment of currency and the last update of the attribute value). Their major idea is to define the metric in a way that its values reflect the context-specific utility resulting from the currency of a considered attribute value. To rescale the

parameter *age of attribute value* to the interval [0; 1], the authors provide examples for possible mathematical rescaling formulations which depend on both a particular user's specific characteristics and a particular application context. This approach is a promising step towards a utility-driven assessment of data quality and constitutes an important contribution in this field.

In the literature there are (a few) very interesting contributions regarding the assessment of the data quality dimension currency which can serve as a basis for further research. However, existing approaches do not aim at a metric that has a concrete unit of measurement and is interval scaled, i.e. the values of the metric, as well as any changes in these values, have to be meaningful. For instance, what does a difference between two values of the metric by Hinrichs of 0.25 exactly mean? The latter is necessary to be able to integrate the values of the metric into a decision calculus (for example, founded on decision theory) and thus to support an economic management of data quality. In addition, almost all existing works do not focus on the goal that their metrics can be configured in an objective way (for example, using statistical methods) and that the assessment can be conducted at a high level of automation (for example, using the metadata in a database). Therefore, we seek to develop a novel metric to assess the currency of data and put a special focus on these aspects in the following.

3. Requirements for data quality metrics for currency

In order to develop a well-founded and applicable metric, we derived six requirements for data quality metrics from literature [5, 8, 14, 15, 32] and they can be summarized as follows.

R.1: Normalization - an adequate normalization is necessary to assure that the values of the metric are comparable (for example, to compare different levels of data quality over time). Because of this, data quality metrics are often ratios with a value between 0 (perfectly bad) and 1 (perfectly good).

R.2: Interval scale - to support both the monitoring of an improved data quality level over time and the economic evaluation of quality measures, we require the metric to be interval scaled. This means that the difference between two values of the metric must be meaningful. For instance, a difference of 0.2 between the values 0.7 and 0.9 and the values 0.4 and 0.6 of the metric shall represent the same extent of improvement of data quality.

R.3: Interpretability - the values of the metric have to be comprehensible and easy to interpret by business users. For instance, considering a metric for completeness, it could be interpretable as the percentage of attribute values which are stored (i.e. they semantically differ from NULL) in the database at the instant of assessment.

R.4: Aggregation - in case of a relational data model, it must be possible to apply the metric on the level of attribute values, tuples, relations (especially views) and the whole database. Here, the values of the metric must have a consistent semantic interpretation on each level. In addition, the metric has to allow aggregation of values on a given level to the next higher level. For instance, the currency of a relation shall be computed based on the currency of the tuples which are part of the relation.

R.5: Adaptivity - to assess data quality in a goal-oriented way, the metric has to be adaptable to the context of a particular application (for example, if the attribute *customer_age* is not relevant in the context of a customer campaign, the metric has to be adapted in order to assign a weight of 0 to this attribute). If the metric is not adapted, it shall fold back to the non-adapted (impartial) assessment.

R.6: Feasibility - to ensure applicability, the metric shall be based on input parameters that are determinable. When defining the metric, methods to determine its input parameters have to be defined. If exact determination is not possible or too cost-intensive, alternative rigorous methods (for example, statistical methods) shall be proposed. From an economic point of view, it is also required that the assessment of data quality can be conducted at a high level of automation.

These six requirements serve as guidance for the design of our novel probabilistic approach to assess the currency of data discussed in the next section.

4. A probability-based data quality metric for currency

Assuring that the values of our metric are normalized (*R.1*), interval scaled (*R.2*), and interpretable (*R.3*) and to enable an automated assessment of data quality (*R.6*), we suggest an approach founded on probability theory. Our novel idea is to interpret currency as the probability that an attribute value stored in a database still corresponds to the current state of its real world counterpart at the instant when data quality is assessed. Such an interpretation provides the necessary indication, which the dimension currency stands for. Considering an address of a library user stored in a database, for example, currency denotes the probability that this address is still up-to-date and has not become outdated. This interpretation of currency is advantageous, because attribute values like addresses 'grow older' and are usually characterized by a shelf life that is unknown by the library. That means that it is unknown how long an address really stays up-to-date, for example. We therefore define a probability-based metric in the following way.

According to the definition of currency given above, we have to assess whether an attribute value (represented by ω) of an attribute (represented by A) stored in a database is still up-to-date.[2] Here, $age(\omega, A) \in \mathrm{IR}^+$ denotes the age of the attribute value, which represents the difference between the instant when data quality is assessed and the instant of data acquisition. In addition, $T \in [0; \infty]$ represents the shelf life of the attribute value, which is usually finite and unknown.[3] We then consider the shelf life to be a continuous random variable and assume that it is exponentially distributed. The exponential distribution is a typical probability distribution for lifetime, which has proven its usefulness in quality management. Moreover, $decline(A)$ represents the average decline rate of the shelf life of attribute values of the attribute under consideration. This parameter can be determined statistically (see Section 5) and represents how many attribute values of the attribute become out of date on average within one period of time. For instance, an average decline rate of 0.2 (i.e. $decline(A) = 0.2$) can be interpreted as follows: on average 20 percent of the attribute's values lose their validity within one period of time. Based on this, we define the metric for currency $Q_{Curr.}(\omega, A)$ as follows:

$$Q_{Curr.}(\omega, A) := \exp(-decline(A) \cdot age(\omega, A)) \tag{3}$$

According to this definition, our metric for currency denotes the probability that the attribute value considered is still up-to-date at the instant when data quality is assessed. For instance, the metric is equal to 1, if an attribute value is acquired at the instant of assessing data quality ($age(\omega, A) = 0 \Rightarrow Q_{Curr.}(\omega, A) = \exp(-decline(A) \cdot 0) = 1$). Hence, in this case it is certain (probability of 100 percent) that the attribute value stored in the database is still up-to-date. Here, the re-collection of an attribute value is considered as a repeated acquisition of an existing attribute value.

Below, we briefly illustrate our metric using a simple example of a library. In this context, it is indispensible to store the borrowers' addresses in a database, for example, to be able to contact them in case of notifications, reminders, problems, etc. Therefore, the attribute *address* seems to be the first choice for applying the metric for currency. To configure the metric defined in term (1) for the attribute *address*, it is necessary to determine the corresponding average decline rate of the values of this attribute ($decline(address)$). This can easily be done for instance by using empirical data from a federal statistics office (for example, the Federal Statistical Office of Germany) considering the frequency of relocation. The average decline rate of about 0.10 per year indicates that on average about 10 percent of the addresses stored in the database become out of date within one year. The resulting values of the metric for currency depending on the age of the attribute value stored are illustrated in Figure 2. It is easy to agree and obvious that the value of the metric, that is the probability that a borrower's address is still up-to-date at the instant when data quality is assessed, decreases over time. For example: the probability for an address, which is one year old, is 90 percent and therefore much higher than the probability for an address acquired 10 years ago. In the latter case, only about 37 out of 100 addresses are still up-to-date.

[2] Please note that there are no restrictions regarding the data type of the attribute values considered. In fact, it is possible to apply the approach for attribute values of type String, Integer, Boolean, etc.

[3] If the shelf life is infinite, the attribute value is always up-to-date. If the shelf life is known, one can decide under certainty whether the attribute value is up-to-date or not. Thus, both cases are trivial or less realistic, respectively.

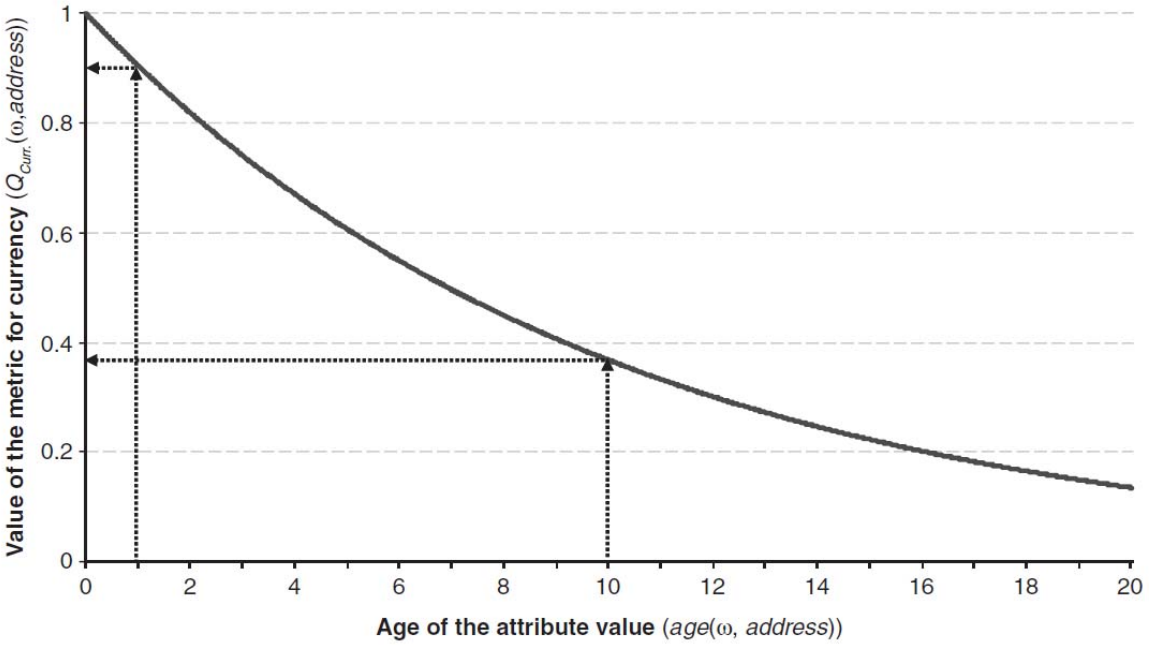

Fig. 2. Value of the metric for currency depending on the age of the attribute value (example)

Term (1) gives the definition of our metric for currency for single attribute values (for example, for the attribute values of the attribute *address*). However, to make it possible to apply the metric to data tuples, relations (especially views), and the whole database as well and therefore to meet the requirement of aggregation (*R.4*), an adequate aggregation has to be defined. In doing so, the metric is constructed 'bottom up'. That means, our metric on the level $n + 1$ (for example, tuples) is based on the metric on the next lower level n (for example, attribute values). Let τ represent a tuple with a certain number (represented by $|A|$) of attribute values (represented by $\tau.A_1$, …, $\tau.A_{|A|}$) for the corresponding attributes (represented by A_1, …, $A_{|A|}$). Moreover, the relative importance of these attributes with regard to currency is given (represented by the weights $g_i \in [0; 1]$ with $\sum_{i=1}^{|A|} g_i > 0$). Consequently, our metric for currency on the level of tuples is defined as the weighted arithmetical mean of the values of the metric for currency on the level of attribute values. More formally, this fact can be denoted as follows:

$$Q_{Curr.}(\tau) := \frac{\sum_{i=1}^{|A|} Q_{Curr.}(\tau.A_i, A_i) g_i}{\sum_{i=1}^{|A|} g_i} \tag{4}$$

Here, the currency of a tuple is calculated on the basis of the values of the metric for the attribute values included (cf. formula (3)). Using the weighted arithmetical mean goes along with three major advantages:

- It is in line with the existing literature [5, 8, 31] and follows the discussion of the major functional forms for developing data quality metrics provided by Pipino et al. [15].
- It is possible to adapt the metric to the context of a particular application as required by *R.5* (cf. relevance of the attributes *address* versus *first_name* in the case of the mobile services provider in Section 5).
- The metric can be applied to relevant subsets of data by assigning weights equal to 0 to attributes which are not relevant within a particular context of application (for example, attribute *telephone_no* within a mailing campaign).

Nevertheless, other functional forms (for example, min operator, i.e. a tuple, is only as current as the least current of all its attribute values) to aggregate the values of the metric may also be appropriate in some cases.

The metric for currency on the levels of relations (especially views) and the whole database can be defined in a similar way [32]. This way, our metric allows data quality on all levels to be adequately assessed (*R.4*).

Table 2 sums up the results and contains the necessary steps for assessing currency. In addition, it denotes whether the steps have to be done manually or whether they can be automated, and how often they are conducted when assessing currency.

Table 2. Steps for assessing currency

1.	**Selection of the data attributes to be evaluated**
	This manual step has to be done only once for a context of application. Here, one has to decide which attributes are relevant regarding an assessment of currency within a particular context of application. For example, for a marketing campaign, in which students are addressed, the attribute *professional_status* with the attribute value *student* has to be selected.
2.	**Configuration of the metric**
	a. Determination of the weights g_i
	This manual step has to be done only once for a context of application. The weights g_i represent the relative importance of the attributes selected in Step 1. For example, within a mailing campaign, the attribute *telephone_no* is not relevant and has to be assigned a weight equal to 0.
	b. Determination of the average decline rates *decline*(*A*)
	This manual step has to be done only once for an attribute since the decline rate of an attribute does not depend on the particular context of application. Here, the decline rate of the attribute *address* may be determined statistically by using external or internal data on relocation.
3.	**Application of the metric**
	a. Computation of the age of the attribute values *age*(ω, *A*)
	This step has to be done for each attribute value. It can be conducted in an automated way by means of Structured Query Language Data Manipulation Language (SQL DML) statements. The age of an attribute value *age*(ω, *A*) is computed based on the metadata in a database (instant when data quality is assessed and instant of data acquisition).
	b. Computation of the values of the metric $Q_{Curr.}(\omega, A)$ on the level of attribute values
	This step has to be done for each attribute value. In this respect, the values of the metric for currency $Q_{Curr.}(\omega, A)$ can be computed automatically by using formula (3) and the input parameters *decline*(*A*) and *age*(ω, *A*).
	c. Aggregation of the values of the metric $Q_{Curr.}(\omega, A)$
	The values of the metric on the level of attribute values $Q_{Curr.}(\omega, A)$ can be aggregated to the levels of tuples, relations (especially views), and the whole database, respectively, in an automated way (cf. formula (4)). For example, to select the customers to be addressed within a marketing campaign, it may be useful to calculate the value of the metric for each customer (on the level of tuples) and select the TOP 30 percent of customers.

The major advantage of our metric compared to existing approaches is the interpretability of its values (cf. *R.3*). Here, the values of the metric represent the probability that a data attribute is still up-to-date at the instant when data quality is assessed. It also ensures that the values of the metric are normalized (cf. *R.1*) and interval scaled (cf. *R.2*). Moreover, we provide adequate formulas (cf. formula (4)) that allow aggregation of the values of the metric in order to make it possible to assess data quality on the level of tuples, relations, and the whole database, too (cf. *R.4*). The weights, which are part of the aggregation formulas, as well as attribute specific decline rates make it possible to assess data quality in a goal-oriented way and ensure the metric's adaptivity to the context of a particular application (cf. *R.5*). Last but not least, Table 2 points out that the probability-based metric enables an assessment of data quality at a high level of automation (cf. *R.6*). The real-world example in the next section illustrates these statements and highlights the applicability of the novel metric as well as its practical benefit.

5. Application of the metric for currency

In cooperation with a major German mobile services provider, we applied our metric in campaign management. For reasons of confidentiality, all specific figures and data had to be changed and made anonymous. Nevertheless, the procedure and the basic results remain the same.

In the past, data quality problems often prohibited successful customer addressing in mailing campaigns and resulted in low campaign success rates. Within Prepaid2Postpaid campaigns, which address prepaid customers forcing them to switch to a postpaid tariff, these data quality problems seemed to be especially evident. One reason for this is that prepaid contracts do not guarantee customer contact at regular intervals (for example, sending bills). Hence, the mobile services provider cannot easily verify for example, whether these customers' contact data are still up-to-date. In the following, we focus on such a Prepaid2Postpaid campaign. Here, about 189,000 customers with the prepaid tariff *Mobile1000* are offered a switch to the postpaid tariff *Mobile2500*. From the viewpoint of the mobile services provider, *Mobile2500* is more profitable since its contract period is fixed and it guarantees minimum sales. Owing to the large number of customers that should be addressed, a real world test to verify each address before mailing the offer out would have been too time-consuming and cost-intensive. Thus, the probability-based novel approach had to show its practical benefit.

First (cf. Table 2), the relevant attributes within the campaign had to be determined. Here, the attributes *surname*, *first_name* and *address* (*street*, *house_number*, *postal_code* and *city* in detail) were considered as relevant for delivering the offer to the customer. Moreover, the customer's current tariff was essential, since it was the selection criterion within the campaign.

Then, the weights g_i of these attributes had to be determined according to their importance within the Prepaid2Postpaid campaign (cf. Step 2a). Since only those customers with the tariff *Mobile1000* should be addressed, the attribute *current_tariff* was stated as most important. Therefore, it was assigned with a weight of 1.0, which served as a reference for the weights of the other attributes. The attribute *address* was considered to be the second most important factor, since without an up-to-date address, the offer cannot be delivered to the customer. Nevertheless, *address* was not given a weight of 1.00, but rather a weight of 0.85, since parts of the address – for instance an up-to-date house number – are not indispensable for the offer's delivery. The attribute *surname* was weighted 0.15 for the following reasons: this attribute is important for the delivery to some extent but, if the surname of a customer changes (after marriage, for example) the old surname might – in many cases – still be known to the postal service (according to experience of the mobile services provider). In contrast, the attribute *first_name* was considered less important. A wrong first name might annoy the customer, but it does not generally prevent delivery. Since the mobile services provider did not want to affect existing customer relationships, *first_name* was assigned the weight 0.05. In the project, the relative weights were set together with the marketing department (alternatively, the importance of the attributes could also have been analysed by means of samples). Furthermore, the average decline rate $decline(A_i)$ had to be determined for each of the selected attributes (cf. Step 2b). Regarding surname and address, we used empirical data from the Federal Statistical Office of Germany on marriages/divorces and the frequency of relocation, respectively. Here, we determined decline rates of 0.02 for the attribute surname (i.e. 2 percent of all customers change their surname per annum) and 0.10 for address. If no third party data had been available, these decline rates could have been estimated by means of internal (historical) data and samples [14]. The decline rate of the attribute *first_name* was assumed as 0.00 since the first name usually remains the same.[4] In contrast, the decline rate of *current_tariff* was estimated based on historical data of the mobile services provider as 0.40 (instead, key account managers could also have been surveyed, accordingly [14]). On this basis, we defined the metric for currency (cf. formulas (3) and (4)).

After the determination of the metric, data quality could be assessed at a high level of automation. For that purpose, the age of each attribute value $age(\tau.A_i, A_i)$ had to be computed automatically (cf. Step 3a). This was done by using SQL DML statements as the instant when data quality is assessed as well as the instant of data acquisition were both available as metadata in the mobile services provider's database. Table 3 shows the

[4] Certainly, the decline rate of the attribute *first_name* is marginal greater than 0. But within the paper, we rounded the decline rates to two digits.

computation of the values of the metric on the level of attribute values for two exemplary customers A and B and the corresponding data tuples (represented by τ_A and τ_B) (cf. Step 3b). Table 3 contains the configuration parameters of the metric which hold for all customers and the customer-specific parameters and computations.

Table 3. Computation of the values of the metric for currency (example for two customer tuples)

A_i		***surname***	***first_name***	***address***	***current_tariff***
g_i		0.15	0.05	0.85	1.00
decline(A_i) [1/year]		0.02	0.00	0.10	0.40
Customer A	$\tau_A.A_i$	'Meier'	'Hans'	'Alstertor 1, 20095 Hamburg'	'Mobile1000'
	age($\tau_A.A_i$, A_i) [year]	0.5	0.5	0.5	0.5
	$Q_{Curr.}(\tau_A.A_i, A_i)$	**0.99**	**1.00**	**0.95**	**0.82**
Customer B	$\tau_B.A_i$	'Huber'	'Peter'	'Mainkai 15, 60311 Frankfurt'	'Mobile1000'
	age($\tau_B.A_i$, A_i) [year]	8.0	8.0	10.0	8.0
	$Q_{Curr.}(\tau_B.A_i, A_i)$	**0.85**	**1.00**	**0.37**	**0.04**

On this basis, it was possible to compute the values of the metric on the level of tuples via aggregation considering the weights (cf. formula (4) and Step 3c):

$$\text{Customer A: } Q_{Curr.}(\tau_A) = \frac{0.99 \cdot 0.15 + 1.00 \cdot 0.05 + 0.95 \cdot 0.85 + 0.82 \cdot 1.00}{0.15 + 0.05 + 0.85 + 1.00} \approx 0.89 \quad (5)$$

For customer B we computed a value of the metric ($Q_{Curr.}(\tau_B)$) of about 0.26 in the same way. Hence, the resulting value of the metric for currency was much higher for the tuple representing customer A than for tuple representing customer B. This means that the data of customer A were much more up-to-date for the context of the application considered here (conducting the marketing campaign). Before applying the metric to the current campaign, we analysed a similar campaign which was conducted three months earlier. Within this campaign 82,000 customers were also offered the chance to switch tariff. The ex-post average success rate was about 8.5 percent; this means that about 7000 customers actually switched. We computed the values of the metric for all customers addressed within this former campaign. Afterwards, we classified the customers according to their values of the metric and assigned each of them to one of the intervals [0; 0.1],]0.1; 0.2], …,]0.9; 1]. Then, we determined the percentage of customers that accepted the offer (campaign success rate) for each interval, accordingly (Figure 3).

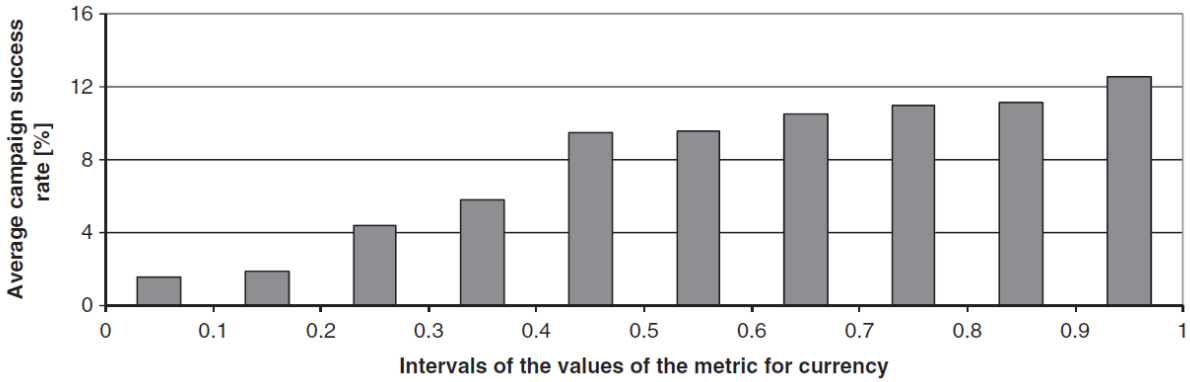

Fig. 3. Success rates of the former campaign depending on the values of the metric.

Figure 3 illustrates that the more up-to-date the attribute values of customers are, the higher the average success rate of the campaign. For instance, the average success rate for the customers within the interval]0.2; 0.3] was only 4.4%, whereas in the interval]0.9; 1] it was 12.6%. This is not surprising, since customers with outdated address data, for example, do not even have the possibility to accept the offer (because they do not even receive it).

But these results became even more interesting when the current campaign is studied. Its target group consisted of 189,000 customers (all customers with tariff *Mobile1000*). Assuming the success rates shown in Figure 3 for this campaign (as the campaign was very similar), the expected additional profit of €20 per customer accepting the offer, and the mailing costs of €1.25 per customer, we were able to calculate the expected profit per customer depending on the customer-specific value of the metric. Here, it became obvious that it does not make sense – from an economic point of view – to address customers with a value of the metric below 0.3. For instance, for the 15,800 customers within the interval]0.2; 0.3] the mailing costs were higher than the expected profit resulting from tariff switching: 4.4% · €20 - €1.25 = €-0.37. The profit from tariff switching did not outweigh the mailing costs (cf. profit (without buying addresses)) until the interval]0.4; 0.5], as depicted in the right chart of Figure 4. By addressing only those customers with a value of the metric higher than 0.4, the profitability of the campaign could be increased by about 30 percent. This is due to the fact that the mailing costs of the campaign are reduced and the expected average success rate rises.

However, an economic management of data quality must not stop at this point. In fact, it is dissatisfying that customers who might switch their tariff are not able to accept the offer, because it cannot be delivered to them (due to outdated address data). Therefore, we analysed if buying external data (as a data quality measure) can help to alleviate the drawback. Firms like the German Postal Services offer up-to-date addresses. Hence, we wanted to find out whether this measure should be taken from an economic point of view. On the one hand, buying address data leads to a cost; on the other hand it improves data quality and may therefore increase campaign success rates. To reach a decision, we firstly calculated the increase of the values of the metric in the case of buying address data for the customers of each interval. This could be done in an automated way via formula (4). On this basis, we were able to estimate the improved campaign success rates by again assuming the success rates of the former, similar campaign depending on the values of the metric. The results are depicted in the left chart of Figure 4. The expected additional profits resulting from higher success rates due to improved data quality were then compared to the costs caused by the purchase of the address data. Here, the costs could be determined easily, since the German Postal Services charged a fixed price of €0.40 per customer address.

6. Findings of the application

The profits made without buying the addresses and the additional profits made by buying external data (expected additional profit resulting from higher success rates due to improved data quality minus costs for buying addresses) are depicted in the right chart of Figure 4.

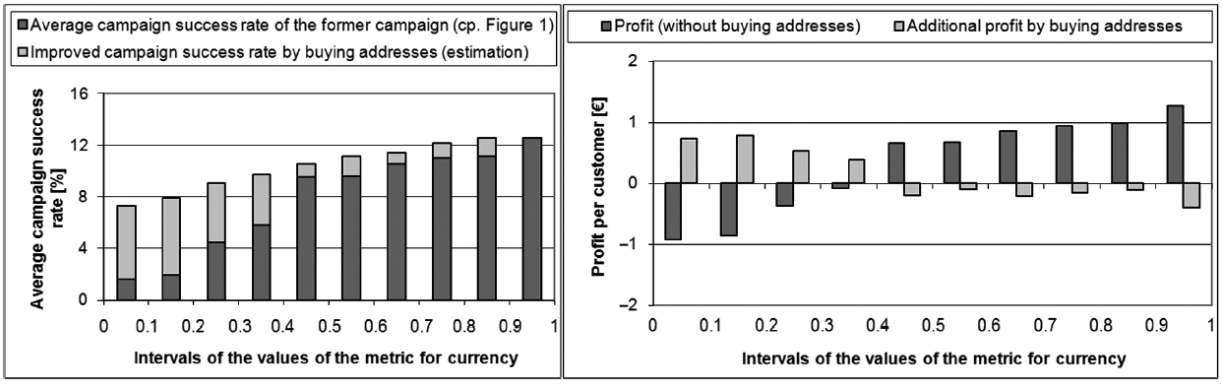

Fig. 4. Campaign success rates and profits depending on the value of the metric.

The results of the analysis are as follows. Without buying addresses it only makes sense to address customers within the interval]0.4; 1] from an economic point of view (positive profits without buying addresses per customer). For these customers buying addresses is not advisable since the costs are higher than the expected additional profits resulting from higher success rates due to improved data quality (negative additional profits by buying addresses per customer). On the other hand, if we do not consider buying addresses at all, it does not make sense to address customers of the interval [0; 0.4] within the campaign (negative profits without buying addresses). However, buying addresses results in positive additional profits for these customers (positive additional profits by buying addresses). For the customers within the interval]0.2; 0.4], the additional profits made by buying addresses are higher than the negative profits without buying addresses – this means that the profits become positive for these customers when buying addresses (profits without buying addresses plus additional profits by buying addresses are greater than 0). In contrast, for the customers within the interval [0; 0.2], the additional profits by buying addresses do not offset the negative profits without buying addresses (profits without buying addresses plus additional profits by buying addresses are smaller than 0).Therefore, following our analysis, buying addresses only makes sense for about 31,900 customers within the interval]0.2; 0.4].

Indeed, the mobile services provider tried to buy addresses for the 31,900 customers characterized by a value of the metric between 0.2 and 0.4 and received data for about 27,500 of them. These addresses were compared to the ones stored in the database. The results of this ex-post analysis are depicted in Table 4.

Table 4. Effect of the data quality measure (ex-post analysis of the values of the metric)

Interval of the values of the metric	**]0.2; 0.3]**	**]0.3; 0.4]**
Number of customers	13,200	14,300
Number of outdated addresses (updated by the measure)	9,600	9,480
Number of valid addresses (not updated by the measure)	3,600	4,820
Fraction of valid addresses ("ex post currency")	**0.273**	**0.337**

The fraction of addresses which were still up-to-date ('ex-post currency') illustrates that the values of our metric for currency (in terms of a probability) calculated ex ante were appropriate (e.g. 0.273 is part of the interval]0.2; 0.3]). In the next step, we recommended sending out the offer to all customers initially assigned to the interval]0.2; 1]. However, as a precaution, the mobile services provider decided to address all 189,000 customers including those characterized by a low value of the metric (interval [0; 0.2]). Ex-post, this turned out to be unprofitable from an economic point of view, because the estimated campaign success rates depending on the value of the metric were quite accurate and the differences between estimated and ex-post campaign success rates were less or equal to ±0.6 percent. For instance, the ex-post campaign success rate of the customers within the interval]0.1; 0.2] was 1.85 percent instead of estimated 1.9 percent and the campaign was (as predicted) not profitable within this interval.

In cooperation with the mobile services provider, the metric was applied in campaign management and led to a substantiated and comprehensible decision support. Specifically, by using the metric the mobile services provider was able to determine a link between the data quality and the success rates of campaigns. Hence, the customer selection process could be improved significantly and costs could also be cut down. Moreover, it was possible to estimate the economic impact of data quality measures in a more accurate way. During the project, most of these improvements could be isolated and analyzed by comparing estimated (ex ante) with realized (ex post) values. Establishing the metric for currency at the mobile services provider was a first, but indispensible, step towards an economic management of data quality.

7. Conclusions

The growing relevance of data quality has revealed the need for adequate measurement. As time aspects are extremely important in data quality management, we propose a novel, probability-based metric for the data quality dimension currency. In contrast to existing approaches, we put a special focus on developing a metric which enables an objective and widely automated assessment of data quality and provides values that are easy to interpret by business users. Therefore, the novel metric meets important requirements for data quality metrics like feasibility and interpretability. The metric on the level of attribute values can be configured in an objective way using statistical methods to derive the average decline rates *decline(A)*. This step does not need to be conducted for each attribute value but only once for each attribute considered. Afterwards, the values of the metric can be computed in an automated way for each attribute value by means of SQL DML statements. Besides, the values of the metric denote the probability that the data considered are still up-to-date. Hence, the metric can be applied to calculate expected values in a methodically well-founded manner in order to support decision making. Precisely, the values of the metric can be integrated into a decision calculus founded on decision theory. In addition, the metric enables an assessment of data quality that does not require a real world test. This is an advantage since comparing attribute values to their real world counterparts is often far too time-consuming and cost-intensive. Moreover, the metric makes it possible to analyse the economic impact of data quality measures adequately. Therefore, it can build the basis for an economic management of data quality.

We demonstrated the metric's practical benefit and its applicability by a real world example. Here, in cooperation with a major mobile services provider, the approach was applied in campaign management (CRM context). Table 5 summarizes the evaluation steps addressed in this paper.

Table 5. Evaluation steps

Evaluation step	Description
Evaluation of the metric with regard to requirements	- Definition of six requirements for data quality metrics derived from literature - Analysis whether the novel metric meets these requirements
Demonstration of the applicability of the metric	- Application of the novel metric in campaign management at a major mobile services provider - Application of the metric to improve the customer selection process and to analyze the economic impact of a data quality measure (buying external data)
Evaluation of the metric values	- An ex post analysis shows that the values of the metric for currency (in terms of a probability) were quite appropriate
Demonstration of the practical benefit of the metric	- An ex post analysis shows that the metric led to an improved decision support - Moreover, applying the metric improved both success rates and profits

The CRM context seems to be especially appropriate to illustrate the practical benefit of the novel metric for currency. Here, the economic effects of data quality and decisions referring to data quality issues can be isolated quite well and quantified in terms of, for example, campaign success rates and profits. This fact makes it easier to calculate business cases and to demonstrate the economic benefit resulting from the application of our approach within this context of application. However, our approach is not limited to the CRM context. In fact, the metric can be applied to data liable to temporal decline (for example, customer data, product data, contract data, employee data, vendor data) and is not restricted to attribute values of certain data types. In fact, it is possible to apply the approach for attribute values of type String, Integer, Boolean, etc. Therefore, the possible fields of application of the novel metric are manifold. Besides the CRM context described above, the metric also approved useful and appropriate in further projects in cooperation with mobile services providers as well as in other fields. For instance, in cooperation with a financial services provider, the metric was effectively applied in a customer valuation project in order to determine expected customer lifetime values [33]. This illustrates that the metric can be used to support decision making in management, production, and logistic processes, if datasets and the expected value serve as a base. Even if the exponential distribution does not hold for specific data attributes (for example, validity of credit or cash card, labour agreement, professional status, marital status, profession, etc.), the metric can be adapted easily by using other types of probability distribution functions. Finally, we can state that the metric fills a gap in both science and practice.

Despite these improvements, there are some limitations of our approach. It is necessary to determine the average decline rates to configure the metric for the attributes considered within a particular context of application. This step can sometimes be time-consuming and cost-intensive. In many cases external data (for example, from federal statistical offices or scientific institutions) can be applied to configure the metric. Otherwise, internal data (for example, from the data warehouse), samples or experts' estimations may be used. However, assessing the data quality dimension correctness by means of a real world test for every single attribute value – which is an alternative to assessing the data quality dimension currency – is usually much more time-consuming and cost-intensive. Moreover, it has to be considered that a metric for currency which was configured once can be reused several times or adapted to other fields of application. Currently, we are working on a model-based economic approach for planning data quality measures. For implementing such a model, adequate data quality metrics are indispensible. The approach presented here provides a basis for these purposes. Nevertheless, further metrics for other data quality dimensions should be developed and further research in this area is strongly encouraged.

References

[1] D.P. Ballou and G.K. Tayi, Enhancing data quality in data warehouse environments, *Communications of the ACM* 42(1) (1999) 73-78.

[2] C.W. Fisher, I.N. Smith and D.P. Ballou, The impact of experience and time on the use of data quality information in decision making, *Information Systems Research* 14(2) (2003) 170-188.

[3] Y.M. Wang and Y.S. Wang, Examining the dimensionality and measurement of user-perceived knowledge and information quality in the KMS context, *Journal of Information Science* 35(1) (2009) 94–109.

[4] L. Al-Hakim, Information quality factors affecting innovation process, *International Journal of Information Quality* 1(2) (2007) 162-176.

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

[6] Harris Interactive, *Information Workers Beware: Your Business Data Can't Be Trusted* (Paris, 2006).

[7] R.Y. Wang, V.C. Storey and C.P. Firth, A framework for analysis of data quality research, *IEEE Transactions on Knowledge and Data Engineering* 7(4) (1995) 623-640.

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

[9] L.P. English, *Improving Data Warehouse and Business Information Quality: Methods for Reducing Costs and Increasing Profits* (Wiley, New York, 1999).

[10] M. Helfert and C. Herrmann, Introducing data quality management in data warehousing. In: R.Y. Wang, E.M. Pierce, S.E. Madnick and C.W. Fisher (eds), *Information Quality, A Volume in the Advances in Management Information Systems*, (M.E. Sharpe, New York, 2005) 135-150.

[11] Z. Jiang, S. Sarkar, P. De and D. Dey, A framework for reconciling attribute values from multiple data sources, *Management Science* 53(10) (2007) 1946-1963.

[12] P. Russom, *Taking Data Quality to the Enterprise through Data Governance* (The Data Warehousing Institute, Seattle, 2006).

[13] M. Pinto, Data representation factors and dimensions from the quality function deployment (QFD) perspective, *Journal of Information Science* 32(2) (2006) 116-130.

[14] B. Heinrich, M. Kaiser and M. Klier, A procedure to develop metrics for currency and its application in CRM, *ACM Journal of Data and Information Quality* 1(1) (2009) 5:1-5:28.

[15] L.L. Pipino, Y.W. Lee and R.Y. Wang, Data quality assessment, *Communications of the ACM* 45(4) (2002) 211-218.

[16] B.D. Klein and T.J. Callahan, A comparison of information technology professionals' and data consumers' perceptions of the importance of the dimensions of information quality, *International Journal of Information Quality* 1(4) (2007) 392-411.

[17] Z. Yu and Y. Wang, An empirical research on non-technical factors related to statistical data quality in China's enterprises, *International Journal of Information Quality* 1(2) (2007) 193-208.

[18] M. Helfert and B. Heinrich, Analyzing data quality investments in CRM – a model based approach. In: *Proceedings of the 8th International Conference on Information Quality (ICIQ)* (University of Cambridge, 2003).

[19] J.M. Juran, How to think about quality. In: J.M. Juran and A.B. Godfrey (eds), *Juran's Quality Handbook* (McGraw-Hill, New York, 1998) 2.1-2.18.

[20] J. Teboul, *Managing Quality Dynamics* (Prentice Hall, New York, 1991).

[21] T.C. Redman, *Data Quality for the Information Age* (Artech House, Boston, 1996).

[22] C. Batini and M. Scannapieco, *Data Quality. Concepts, Methodologies and Techniques* (Springer, Berlin, 2006).

[23] M.J. Eppler, *Managing Information Quality* (Springer, Berlin, 2003).

[24] M. Jarke and Y. Vassiliou, Foundations of data warehouse quality – a review of the DWQ project. In: *Proceedings of the 2nd International Conference on Information Quality (ICIQ)*, (University of Cambridge, 1997).

[25] Y.W. Lee, D.M. Strong, B.K. Kahn and R.Y. Wang, AIMQ: a methodology for information quality assessment, *Information and Management* 40(2) (2002) 133-146.

[26] C. Cappiello, C. Francalanci and B. Pernici, Time-related factors of data quality in multichannel information systems, *Journal of Management Information Systems* 20(3) (2003) 71-91.

[27] R.R. Nelson, P.A. Todd and B.H. Wixom, Antecedents of information and system quality: an empirical examination within the context of data warehousing, *Journal of Management Information Systems* 21(4) (2005) 199-235.

[28] D.P. Ballou and H.L. Pazer, Modeling data and process quality in multi-input, multi-output information systems, *Management Science* 31(2) (1985) 150-162.

[29] B. Otto, K.M. Hüner and H. Österle, Identification of business oriented data quality metrics. In: *Proceedings of the 14th International Conference on Information Quality (ICIQ)* (HPI Potsdam, 2009).

[30] A. Even and M. Kaiser, A Framework for Economics-Driven Assessment of Data Quality Decisions. In: *Proceedings of the 15th Americas Conference on Information Systems (AMCIS)* (Association for Information Systems, 2009).

[31] H. Hinrichs, *Data quality management in data warehouse systems. PhD thesis* (University of Oldenburg, 2002) (in German).

[32] B. Heinrich, M. Kaiser and M. Klier, How to measure data quality? – a metric based approach. In: *Proceedings of the 28th International Conference on Information Systems (ICIS)* (Association for Information Systems, 2007).

[33] B. Heinrich and M. Klier, A novel data quality metric for timeliness considering supplemental data. In: *Proceedings of the 17th European Conference on Information Systems (ECIS)* (University of Verona, 2009).