University of Augsburg
Prof. Dr. Hans Ulrich Buhl
Research Center
Finance & Information Management
Department of Information Systems
Engineering & Financial Management

Universität Augsburg
University

Discussion Paper WI-361

An Indicator Function for Insufficient Data Quality - A Contribution to Data Accuracy

by

Quirin Görz, Marcus Kaiser

March 2012

in: Lecture Notes in Business Information Processing, 7th Mediterranean Conference on Information Systems, Guimarães, Portugal, September 2012, 129, p. 169-184.

Universität Augsburg, D-86135 Augsburg
Visitors: Universitätsstr. 12, 86159 Augsburg
Phone: +49 821 598-4801 (Fax: -4899)
www.fim-online.eu

An Indicator Function for Insufficient Data Quality – A Contribution to Data Accuracy

Quirin Görz [1] and Marcus Kaiser[2],

[1] FIM Research Center, University of Augsburg, Universitätsstr. 12, 86135 Augsburg, Germany
quirin.goerz@wiwi.uni-augsburg.de

[2] Senacor Technologies AG, Erika-Mann-Str. 55, 80636 Munich, Germany
marcus.kaiser@senacor.com

Abstract. Owing to the fact that insufficient data quality usually leads to wrong decisions and high costs, managing data quality is a prerequisite for the successful execution of business and decision processes. An economics-driven management of data quality is in need of efficient measurement procedures, which allow for a predominantly automated identification of poor data quality. Against this background the paper investigates how metrics for the DQ dimensions completeness, validity, and currency can be aggregated to derive an indicator for accuracy. Therefore existing approaches to measure these dimensions are analyzed in order to make explicit, which metric addresses which aspect of data quality. Based on this analysis, an indicator function is designed returning a measure for accuracy on different levels of a data resource. The indicator function's applicability is demonstrated using a customer database example.

Keywords: Data quality, data quality management, measurement, accuracy.

1 Introduction

Poor data quality (DQ) usually is associated with wrong decisions and high costs [3], [16], [19]. For instance, a study conducted by the Data Warehouse Institute revealed that in 67% of the involved organizations poor DQ causes high costs [38]. Moreover, 75% of the interviewees in an international study on DQ admitted wrong decisions due to incorrect data [23]. These findings are complemented by a survey conducted by CSO Insights, where 47% of the responding senior marketing executives had seen either a noticeable or severe impact on their marketing campaigns from poor DQ [13]. Consequently, high quality data are prerequisite for executing business and decision processes. Ensuring DQ hence constitutes a relevant problem for organizations [4], [38]. To solve this problem in an economics-oriented manner, efficient instruments are necessary to measure and subsequently detect insufficient DQ [22], [24], [37].

Literature indicates that DQ is a multidimensional concept [35], [42]. One of the most cited DQ dimension is accuracy [32], [40]. To measure accuracy of an attribute

value exactly, two pieces of information must be available: The attribute value as it is stored in a data resource and its real-world counterpart. In many cases, determining the latter is time- and cost-intensive [20]. Thus, there is a need for other DQ dimensions which allow for deriving information on an attribute value's quality without knowing its (costly to determine) real-world counterpart. DQ dimensions which can be used to indicate an attribute value's accuracy are completeness, validity, and currency. However, these DQ dimensions only give an indication on an attribute value's accuracy and do not represent its exact accuracy. Moreover, each of these dimensions only represents one particular aspect of the accuracy of an attribute value. Thus, an aggregated view of these three dimensions would yield a more complete indication of an attribute value's accuracy.

Although it is well known that there are interdependencies between particular DQ dimensions [2], [10], DQ dimensions and especially their measurement are usually discussed independently from each other. So far, three approaches exist to aggregate metrics for different DQ dimensions, which will be discussed in this paper. This discussion will reveal that there is no approach to estimate in a consistent, meaningful way, how accurate data stored in a data resource are, even if the attribute values are already evaluated via metrics for different DQ dimensions. This is the more astonishing, as many papers in the area of DQ management rely on a variable representing the overall level of DQ without defining this variable in a formal way (e.g. [18]). Within this paper we aim to close this research gap by designing an indicator function for accuracy. Therefore we investigate the following research question: *How can accuracy of an attribute value be measured by aggregating metrics for the DQ dimensions completeness, validity, and currency?*

The paper is organized as follows: Section 2 sums up the relevant literature dealing with defining and measuring the DQ dimensions accuracy, completeness, validity, and currency as well as on approaches for aggregating metrics for different DQ dimensions. Afterwards, an indicator function for indicating accuracy is designed in Section 3. In Section 4 the indicator function's applicability is demonstrated by means of a customer database example. Results and limitations are summarized in Section 5.

2 Background and Related Work

Literature provides several definitions for DQ. For instance, according to Orr [34] DQ "is the measure of the agreement between the data views presented by an information system and that same data in the real-world" and Parssian et al. [36] state that "the terms information quality and data quality have been used to characterize mismatches between the view of the world provided by an IS and the true state of the world". To characterize these "mismatches" in more detail, several DQ dimensions have been introduced. According to Lee et al. [32], the dimensions accuracy, completeness, validity, and currency are most relevant in the context of measuring the quality of attribute values. One important feature of these four dimensions is, that they refer to a

quality of conformance (QoC) perspective[1] and hence can be measured by metrics in an inter-subjectively verifiable manner [30]. In the following we discuss definitions and metrics for these four dimensions, relate them to each other with respect to their measurement costs, and give a brief overview on several approaches to aggregate DQ dimensions. From this discussion, we deduce the research gap to be investigated in this paper.

2.1 Definitions and measurement procedures

To get an overview which aspects of DQ the four DQ dimensions accuracy, completeness, validity, and currency cover and how these aspects can be measured by means of metrics, we discuss their definitions and the respective measurement procedures in the following. Thereby we focus on metrics, which have been formally defined and can be measured in a predominantly automated way. Other well-known approaches (for an overview, please refer to Batini et al. [5]) like the AIMQ-Method [32] or the Total Data Quality Methodology [41] are hence not in the scope of this paper as they do not provide formally defined metrics which allow for a predominantly automated measurement.

Accuracy. According to Eppler [15], 'accuracy' can be defined as "how closely information matches a real-life state", which is a common definition in DQ literature and will be used in the following. That is, accuracy of an attribute value measures the distance between the attribute value $v_I(t_n.a_m)$ stored in the data resource and the corresponding value in the real-world $v_W(t_n.a_m)$ at the moment of measuring DQ [6]. If these two values are identical, the stored attribute value $v_I(t_n.a_m)$ is correct. Thereby, t_n ($n \in IN$) stands for a tuple which represents a real-world entity by storing its attribute values $t_n.a_1$, $t_n.a_2$, ... $t_n.a_M$. Accuracy is thus determined by means of a distance measure [20]:

Let $d(v_I(t_n.a_m), v_W(t_n.a_m))$ denote a domain-specific distance function for attribute a_m. It quantifies the closeness between $v_I(t_n.a_m)$ and $v_W(t_n.a_m)$ and normalizes its results to the interval [0; 1], where 0 connotes perfect congruence and 1 connotes no congruence at all[2]. Based on these definitions we can define the metric for accuracy on the level of attribute values as follows [25]:

$$Q_{Accu}(v_I(t_n.a_m), v_W(t_n.a_m)) = 1 - d(v_I(t_n.a_m), v_W(t_n.a_m)) \quad (\mathbf{1})$$

So, to measure accuracy, not only the attribute value $v_I(t_n.a_m)$ stored in a data resource is needed, but also its counterpart in the real-world $v_W(t_n.a_m)$ at the moment of measurement. This necessity of having to know the real-world counterpart $v_W(t_n.a_m)$ is a shortcoming in practical application [20], which can only be solved at high cost in certain areas – if at all [24]. Taking the example of customer data, each customer has to be contacted and asked for the correct value which holds at the moment of asking. Although this is an extreme example, acquiring the real-world counterpart nearly

[1] More details on the two perspectives "quality of conformance" (QoC) and "quality of design" (QoD) can be found in Heinrich et al. [25] and Juran [29].

[2] Please refer to Heinrich et al. [26] for examples of normalized distance functions and illustrations for the effects of not normalized distance functions.

always occasions high costs. On the other hand, after such a comparison of the attribute value stored $v_I(t_n.a_m)$ to its real-world counterpart $v_W(t_n.a_m)$, the latter is known and as a result, $v_I(t_n.a_m)$ can be updated and should therefore be perfectly accurate.

Completeness. Literature on DQ uses the term ‘completeness‘ in different contexts. As this paper deals with the quality of attribute values in a data resource which is assumed as given (cf. QoD perspective), an attribute shall be defined as complete in this paper (in accordance with Fox et al. [20] and Batini and Scannapieco [6]), if it semantically differs from *NULL*. *NULL* is equivalent to “missing value” [39] and means “value at present unknown” [12]. This is in contrast to Fox et al. [20], who – in addition – perceive “value does not exist” as *NULL*. If it is however known, that no value exists for an entity’s attribute (e.g. a customer has no telephone and therefore no value can be stored for the attribute ’phone number’), this attribute should be considered as complete in terms of the definition above. Hence, it is represented by a corresponding (standardised) attribute value in the data resource (e.g., ’N/A’ for “property inapplicable” [12] or “nothing” [39]).

To measure the completeness of an attribute value, one has to determine a set of attribute values $S_{Incomp}^{a_m}$ for each attribute a_m, considered as incomplete. The elements in this set are thus semantically equivalent to *NULL*. Once $S_{Incomp}^{a_m}$ is determined, a metric for completeness (slightly adapted from the metric defined by Heinrich et al. [25]) of an attribute value $v_I(t_n.a_m)$ can be defined:

$$Q_{Comp}(v_I(t_n, a_m)) = 0 \Leftrightarrow v_I(t_n, a_m) \in S_{Incomp}^{a_m} \quad \textbf{(2a)}$$

$$Q_{Comp}(v_I(t_n, a_m)) = 1 \Leftrightarrow v_I(t_n, a_m) \notin S_{Incomp}^{a_m} \quad \textbf{(2b)}$$

In comparison to measuring accuracy it is not necessary to determine the real-world counterpart of the attribute value $v_I(t_n.a_m)$ stored in the data resource. Instead it is sufficient to know the attribute value $v_I(t_n.a_m)$ and the set of attribute values $S_{Incomp}^{a_m}$, which are considered as incomplete. Thus, it is possible to measure completeness repeatedly and – at least, to a large extent – automatically. This is why it should cause less effort to measure completeness than accuracy.

Validity. An attribute value “is invalid if its contents are not within the pre-specified value domain […] and is valid otherwise” [17], a definition which shall also be used in this article. Moreover, ‘validity’ is also known as ‘domain integrity’ which means that “all values of an attribute must be drawn from a specified domain” [31]. Hence, from a QoC perspective, validity refers to the question, whether an attribute value $v_I(t_n.a_m)$ is part of a (pre-specified) value domain $S_{Vali}^{a_m}$ or not. As with the set of incomplete values, defining the value domain $S_{Vali}^{a_m}$ for the particular attribute a_m is a prerequisite for measuring its validity. The value domain $S_{Vali}^{a_m}$ can also be defined by means of several rules or constraints [27]. Moreover, the validity of an attribute value

is measured by means of a Boolean variable [20]. According to Heinrich et al. [27], an attribute value is thus either *valid* or *not valid*. With $S_{Vali}^{a_m}$ being the value domain, that is, the set of all values which are valid for attribute a_m, we define the metric for validity on the level of attribute values as follows:

$$Q_{Vali}(v_I(t_n, a_m)) = 0 \Leftrightarrow v_I(t_n, a_m) \notin S_{Vali}^{a_m} \quad \textbf{(3a)}$$

$$Q_{Vali}(v_I(t_n, a_m)) = 1 \Leftrightarrow v_I(t_n, a_m) \in S_{Vali}^{a_m} \quad \textbf{(3b)}$$

Thus, the value domain $S_{Vali}^{a_m}$ can be deduced from business rules or domain-specific functions [27]. The valid value domain of an attribute value $v_I(t_n.a_m)$ can also be derived from another attribute value $v_I(t_n.a_o)$ for attribute a_o of the same tupel t_n (we use the term 'tupel' instead of 'record set' to avoid confusion with respect to the variables, as we will use *r* for 'relation' later on). However, if such cross-attribute logical dependencies shall be taken into account, it is necessary to analyze the validity of the determining attribute value $v_I(t_n.a_o)$ first. This is because restricting the value domain for the dependent attribute value $v_I(t_n.a_m)$ based on the determining value $v_I(t_n.a_o)$ for another attribute is only reasonable if the determining value $v_I(t_n.a_o)$ is valid itself. Taking the example of an address, it does not make sense to derive the 'city' from the 'zip code', if the latter value is for instance negative, i. e. not valid. Owing to this restriction, we leave aside cross-attribute logical dependencies for this paper, but such effects are subject to further research.

Completeness and validity have in common, that a set of values has to be defined in advance so that the two dimensions can be measured. In general, the effort for defining the corresponding set should be higher for validity, as all valid values have to be defined for a particular attribute, whereas completeness requires only listing of those values which are semantically equivalent to *NULL* for each attribute. Comparing the effort for measuring validity to the one for accuracy, the effort for the former can be considered by far lower, as it does not require a real-world test and – after the initial definition of $S_{Vali}^{a_m}$ – the measurement can be repeated in an automated way.

Currency. Recent papers on DQ define 'currency' (often used synonymously with timeliness) as the probability that an attribute value $v_I(t_n.a_m)$, which was accurate at the instance of its storage, is still congruent with its real-world counterpart $v_W(t_n.a_m)$ at the moment of measurement [24]. This definition shall also hold for this article. That is, currency represents the probability that an attribute value is still up-to-date and has not become outdated due to a temporal decline. Thus, in contrast to measuring accuracy, measuring currency provides a probability and not a verified statement under certainty.

Let hence $s^{v_I(t_n.a_m)}$ be the time period which has passed since the accurate storage of the attribute value $v_I(t_n.a_m)$ in the data resource. Furthermore, S^{a_m} denotes the shelf life of an attribute a_m. That is, it represents the time period which passes before the – originally accurately stored – attribute value $v_I(t_n.a_m)$ does no longer correspond to its

real-world counterpart, as $v_W(t_n.a_m)$ has changed in the meantime. As shelf life S^{a_m} is usually unknown, it is considered a random variable. Moreover, according to Heinrich et al. [24], we define a distribution function $F^{a_m}\left(s^{v_I(t_n.a_m)}\right) = P\left(S^{a_m} \leq s^{v_I(t_n.a_m)}\right)$ specifically for each attribute a_m. This distribution function returns the probability that the shelf life of a particular attribute value is shorter than the time period which has passed since its storage in the data resource; or – to put it another way – the probability that an attribute value became outdated in the meantime due to a temporal decline. Based on these definitions the general metric for currency can be defined as follows [24]:

$$Q_{Curr}(s^{v_I(t_n,a_m)}) = 1 - F^{a_m}(s^{v_I(t_n,a_m)}) \quad (4)$$

In contrast to completeness and validity, the metric for currency is not based on a set of values defined for each attribute, but on the distribution function $F^{a_m}\left(s^{v_I(t_n.a_m)}\right)$ which has to be specifically determined for each attribute a_m. To do this, statistical procedures are necessary as, for instance, discussed in Heinrich et al. [24]. This investment has to be made once before the first measurement, but the resulting distribution function can then be used several times. As a result, the recurring costs for measuring currency are usually less than for measuring accuracy.

Summing up, the definition of the DQ dimension accuracy seems to be the closest to the definitions of DQ: By measuring accuracy insufficient DQ is certainly detected and the DQ dimensions completeness, validity, and currency are measured simultaneously. But when measuring accuracy, not only the attribute value $v_I(t_n.a_m)$ stored in a data resource is needed, but also its real-world counterpart $v_W(t_n.a_m)$ at the instance of measurement. In contrast, measuring completeness, validity, and currency can be realized by means of metrics without comparing the stored attribute value $v_I(t_n.a_m)$ to its real-world counterpart $v_W(t_n.a_m)$. Although, determining a set of incomplete attribute values $S^{a_m}_{Incomp}$, a value domain $S^{a_m}_{Vali}$, and a distribution function $F^{a_m}\left(s^{v_I(t_n.a_m)}\right)$ occasions initial costs, measuring accuracy usually is much more expensive in the long run. This is because the parameters for measuring completeness, validity, and currency have to be determined once and can afterwards be used for multiple automated measurements with no or little adaptions, while the real-world counterpart for measuring accuracy has to be determined for each measurement at high cost anew. Owing to these high costs, indicating an attribute value's accuracy, without knowing its real-world counterpart, would be of high value, especially in recurring measurements. Thus, taking into account metrics for completeness, validity, and currency – which are less cost intensive in the long run – seems to be economically reasonable. As each of these metrics measures a specific aspect of an attribute value's accuracy (see below) an aggregated view of these three dimensions would yield a more complete indication of an attribute value's accuracy and hence of its insufficiency. Consequently, approaches to aggregate DQ dimensions are discussed in the following.

2.2 Approaches to aggregate different DQ dimensions

Existing interdependencies between particular DQ dimensions are analysed in several publications. For instance, interdependencies between DQ dimensions like accuracy and timeliness (not currency) as well as completeness and consistency are modelled as trade-offs [1], [2]. In addition, logical connections between DQ dimensions are discussed [21] and approaches are developed to quantify existing interdependencies by means of correlations [14], [32]. Besides, the dependencies and the interactions between different DQ dimension are analysed based on the complexity of the problem [7]. None of these papers addresses the topic of formally aggregating DQ dimensions. So far, only three publications do so:

The first [9] designs a formal approach for combining the results of metrics for different DQ dimensions to one aggregated DQ measure. It is defined based on a weighted average:

$$Q_{Over} = w_{Accu} \text{x} Q_{Accu} + w_{Comp} \text{x} Q_{Comp} + w_{Vali} \text{x} Q_{Vali} + w_{Curr} \text{x} Q_{Curr} + w_{Inter} \text{x} Q_{Inter} + w_{Acc} \text{x} Q_{Acc} \quad (5)$$

This aggregation takes into account the dimensions 'interpretability' (Q_{Inter}) and 'accessibility' (Q_{Acc}), which refer to Quality of Design (QoD) and not to QoC. Hence, these dimensions are not relevant in our context. Nevertheless, the idea of using a weighted average might still be appropriate to aggregate metrics for different DQ dimensions, as also Pipino et al. [37] propose to do so. However, using a weighted average comes along with several shortcomings, which are discussed by Helfert et al. [28]. They mainly stress that a weighted average assumes independence of the metrics to be aggregated. As will be revealed in section 3, this assumption does not hold in the given context: for instance, an incomplete value should not be valid; consequently, independence is not given.

Besides, Even and Shankararayanan [16] suggest an aggregation function which shall reflect the overall utility reduction caused by different quality defects. They propose the algebraic product:

$$Q_{Cons} = Q_{Accu} \text{ x } Q_{Comp} \text{ x } Q_{Vali} \text{ x } Q_{Curr} \quad (6)$$

On the one hand, this aggregated quality of an attribute value is perfect ($Q_{Cons} = 1$) if no defect is present (i.e. $Q_{Accu} = Q_{Comp} = Q_{Vali} = Q_{Curr} = 1$) and on the other hand, it is absolutely imperfect ($Q_{Cons} = 0$) if at least one of the components has a zero value. However, within the paper of Even and Shankararayanan [16], completeness refers to a QoD and not to a QoC definition: It is defined as the inclusion or exclusion of an attribute value in the data specification. This makes sense from a utility based point of view, but not for indicating an attribute value's accuracy.

In addition, Calero et al. [8] develop a DQ model for web portals (the so called PDQM). That is, they develop a model to determine overall DQ of a web portal based on probabilistic theory. Therefore, Calero et al. [8] use an approach that employs Bayesian networks and Fuzzy logic in order to aggregate several DQ dimensions. These DQ dimensions mostly rely on a QoD definition (e.g. applicability, availability, believability, flexibility, etc.) rather than a QoC definition. Furthermore, they also do not measure accuracy by a real-world test but give a discrete indication of accuracy (good, medium, bad) based on the number of duplicates presented on a web portal.

The approach is feasible for determining overall DQ of web portals and has been partly tested in a real-world setting [11] but is insufficient for other domains such as corporate data sets as several DQ dimension (e.g. accuracy, completeness, etc.) have been adapted to the web portal domain [8].

Moreover, in these approaches another problem arises by including the dimension accuracy itself: As argued earlier, it is necessary to determine the real-world value $v_W(t_n.a_m)$ in order to measure accuracy, so that it can be compared to the attribute value stored in the data resource $v_I(t_n.a_m)$. Thereby, accuracy can be determined exactly – whereas the other three dimensions only return an indication on the accuracy of the attribute value. If the real-world value is known, the metric for accuracy should be used and there is no need for the other dimensions or their aggregation at all.

2.3 Research gap

So far, there exists no approach to aggregate metrics for completeness, validity, and currency enabling an indication on the accuracy of an attribute value by considering dependencies between the different DQ dimensions and representing them adequately. To close this research gap, we design an indication function which makes use of metrics for the DQ dimensions completeness, validity, and currency to indicate an attribute value's accuracy in an economics-oriented way.

3 Design of an Indicator Function for Accuracy

The indicator function is an artefact which is to be designed. To guide the search process for this artefact in a scientifically founded way, we state eight requirements which shall be fulfilled by the indicator function.

First, we demand five requirements which were already used in existing DQ literature to derive metrics to measure the four DQ dimensions considered in this paper: completeness [25], validity [27], currency [24], and accuracy [26]. That is, the metrics to measure these dimensions (represented by formulas 1 to 4) also meet the following requirements:

(R1) Normalisation: The results of the indicator function must be normalised to ensure that they can be compared to each other (e.g. to compare different levels of DQ over time [37]). In this context, DQ metrics are often ratios with a value ranging between 0 (perfectly bad) and 1 (perfectly good) [16], [37].

(R2) Interval scale: The difference between two results of the indicator function must be interval scaled (i.e. must have a defined meaning which remains the same independent from the height of the results). Only then the results can be input parameters to economic considerations.

(R3) Interpretability: Only if the meaning of the results of the indicator function are comprehensible, they are "easy to interpret by business users" as demanded by Even and Shankaranarayanan [16].

(R4) Aggregation: It shall be possible to quantify DQ on the level of attribute values, tupels, relations, and the whole (relational) database in a way, so that the values have consistent semantic interpretation (interpretation consistency, [16]) on each level. In addition, the metrics must allow aggregation of values on a given level to the next higher level (aggregation consistency, [16]).

(R5) Applicability: 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.

Requirements (R1) to (R3) characterise the results of the indicator function, whereas (R4) and (R5) address the applicability in the context of an economics-oriented DQ management.

The further requirements define how the different DQ dimensions considered impact the result of the indicator function on the level of attribute values. Hence, we are looking for a function $Q_{Ind}^{v_I(t_n.a_m)}(Q_{Comp}(v_I(t_n.a_m)), Q_{Vali}(v_I(t_n.a_m)), Q_{Curr}(s^{v_I(t_n,a_m)}))$, which returns an indicator on the accuracy of the attribute value $v_I(t_n.a_m)$ based on its metric results for completeness $Q_{Comp}(v_I(t_n.a_m))$, validity $Q_{Vali}(v_I(t_n.a_m))$, and currency $Q_{Curr}(s^{v_I(t_n,a_m)})$.

Again, we start with completeness: An incomplete attribute value cannot be accurate by definition; if no attribute value is stored, it is different from the real-world counterpart (if the latter exists – cf. section 2.1.2). Consequently, it cannot be valid or current either and it would therefore seem inappropriate if the metric results for validity and currency had any influence. Hence, in case of an incomplete attribute value, the metric for completeness fixes the value of the indicator function at 0:

(R6) $Q_{Comp}(v_I(t_n.a_m)) = 0 \Rightarrow Q_{Ind}^{v_I(t_n.a_m)}(0, Q_{Vali}(v_I(t_n.a_m)), Q_{Curr}(s^{v_I(t_n,a_m)})) = 0$.

Note that the metric for currency might return a value greater than 0, as it relies only on the shelf life $s^{v_I(t_n.a_m)}$ and does not take into account the actually stored value. This indication is however overruled via (R6), so that solely completeness determines the overall value of the indicator function.

If an attribute value is however complete, the result of the indicator function depends on the dimensions validity and currency. As an invalid value is also inaccurate by definition, currency is not relevant here either and it shall hold:

(R7) $Q_{Comp}(v_I(t_n.a_m)) = 1 \wedge Q_{Vali}(v_I(t_n.a_m)) = 0 \Rightarrow Q_{Ind}^{v_I(t_n.a_m)}(1,0, Q_{Curr}(s^{v_I(t_n,a_m)})) = 0$

Again, (R7) ensures that currency is overruled and has no impact on the overall indication.

In case of a complete and valid attribute value, its DQ will be judged in addition based on its currency, because only in this case the attribute value can be accurate and the metric for currency indeed returns the probability that the stored attribute value still corresponds to its real-world counterpart. Consequently, it shall hold:

(R8) $Q_{Comp}(v_I(t_n.a_m)) = 1 \wedge Q_{Vali}(v_I(t_n.a_m)) = 1 \Rightarrow Q_{Ind}^{v_I(t_n.a_m)}(1,1, Q_{Curr}(s^{v_I(t_n,a_m)})) = Q_{Curr}(s^{v_I(t_n,a_m)})$

Comparing the existing approaches to (R1) to (R8) it can be stated that the weighted average operator (5) meets requirements (R6) to (R8) only if the weights are determined for each attribute value $v_I(t_n.a_m)$ individually based on its completeness, validity, and currency: For instance, if an attribute value is complete, but invalid, the weights for completeness and currency should be 0, whereas the weight for validity should be 1. This procedure seems rather complex and causes additional computation time when applying the indicator function in an automated way. Moreover, the purpose of the weights is not to fully exclude or include one dimension, but to provide a weighting based on the dimensions general relevance. Consequently, the weighted average operator seems not suitable for our purposes. Although, the Bayesian network approach proposed by Calero et al. [8] is based on probabilistic theory it cannot be applied to the metrics introduced in section 2. The PDQM relies on specific (objective and subjective) measures for the Bayesian network's entry nods which form the basis for determining the measures of the specific DQ dimensions (e.g. accuracy) in terms of probability tables.

A mathematical operator fulfilling requirements (R1) to (R8) is the algebraic product, which was also used in (6). Based on it, the indicator function for accuracy can be formulated as follows:

$$Q_{Ind}^{v_I(t_n.a_m)}(Q_{Comp}(v_I(t_n.a_m)),Q_{Vali}(v_I(t_n.a_m)),Q_{Curr}(s^{v_I(t_n,a_m)})) = \quad (7)$$
$$= Q_{Comp}(v_I(t_n.a_m)) \times Q_{Vali}(v_I(t_n.a_m)) \times Q_{Curr}(s^{v_I(t_n,a_m)})$$

The value of this indicator function is 0, if the attribute value is incomplete or invalid or both. In all cases, currency is not taken into account. Only if an attribute value is complete and valid, currency plays a role and determines the result of the indicator function.

Besides (R6) to (R8), the proposed indicator function fulfils the other properties as well: On the level of attribute values, the results of formula (7) are normalized to the interval [0; 1] (R1). As only the value domain for currency is the continuum between 0 and 1, the interval scale property depends on this dimension. Since currency is measured by means of a probability, the results are interval scaled (R2). In addition, a probability can be considered interpretable (R3). Once the initial actions for an automated measurement are taken per attribute (definition of the corresponding sets $S_{Incomp}^{a_m}$ and $S_{Vali}^{a_m}$ for completeness and validity or definition of the distribution function $F^{a_m}(s^{v_I(t_n.a_m)})$ for currency respectively), the measurement can be done repeatedly in an automated way (R5).

To meet requirement (R4), we also develop formulas which give an indication on accuracy on higher levels of a data resource based on formula (7). As the metrics have to be defined specifically for each attribute, the attributes of a relation shall be considered as the next level [33] (in contrast to e.g. Heinrich et al. [25], who consider the tupels of a relation as the next level).

The indicator function on the level of attribute a_m bases on the indicator function values of all $N \in IN$ tupels, which are stored in a relation at the moment of measuring

DQ. To measure DQ in an inter-subjectively verifiable way, the tupels t_n shall not be weighted, so that all have the same impact:[3]

$$Q_{Ind}^{a_m} = \sum_{n=1}^{N} Q_{Ind}^{v_I(t_n.a_m)}(Q_{Comp}(v_I(t_n.a_m)), Q_{Vali}(v_I(t_n.a_m)), Q_{Curr}(s^{v_I(t_n,a_m)})) \Big/ N \quad (8)$$

Also on the level of relations, the indicator function for accuracy can be determined based on the indicator function of the (unweighted) $M \in IN$ attributes by means of the arithmetic mean:

$$Q_{Ind}^{r} = \sum_{m=1}^{M} Q^{a_m} \Big/ M \quad (9)$$

Assuming the data resource consisting of $P \in IN$ pairwise non overlapping relations and all attributes being represented only once in the data resource, the indicator function for accuracy on the level of the data resource can be defined using formula (9):

$$Q_{Ind}^{d} = \sum_{r=1}^{P} Q^{r} \Big/ P \quad (10)$$

The designed formulas can now be used to give an indication on the DQ in terms of accuracy on all levels of a data resource in an inter-subjectively verifiable and automated way. Hence, (A4) is met.

4 Demonstration of Applicability

The practical applicability of the indicator function for accuracy shall be demonstrated in a customer database example. We consider a fictive company intending to measure the accuracy of its data at regular intervals. The company preferably conducts e-mail-based direct marketing campaigns. The success of these campaigns depends on the accuracy of the e-mail addresses stored in the customer database. Therefore the company measures at regular intervals the quality of e-mail addresses stored in the customer database. The customer database is built using a relational database schema. For reasons of clearness, we consider one relation "customers", which is shown in Table 1.

Table 1 Exemplary relation for customer data

C_ID	last_name	first_name	e_mail_address	entry_date
1	Hansen	Olaf	O.Hansen@example.com	1998-11-01
2	Parker	Peter	p.parker@world-time.time	2010-01-17
3	Smith	Michael	*NULL*	2007-09-27

[3] Some existing approaches propose to weight attributes and/or tupels. Such a weighting can be useful in particular business situations and can be integrated in the formulas proposed here.

The relation consists of the following five attributes: a distinct identifier (C_ID), a customer's last name (last_name), a customer's first name (first_name), a customer's e-mail address (e_mail_address), and the respective date of data entry (entry_date). To not find out only with the next campaign about the inaccuracy of the e-mail addresses, the company applies the indicator function designed above. At the level of attribute values, the indicator function is designed as a product of the results of the metrics for completeness, validity, and currency. For calculating the indicator function, the company thus has to take the following four steps: (i) calculate the metric for completeness, (ii) calculate the metric for validity, (iii) calculate the metric for currency, and (iv) multiply the metrics' results for each attribute value. All four steps can be performed in a predominantly automated way. This shall be illustrated by describing the measurement of the metrics in terms of the standard data query language SQL (while acknowledging that other ways of implementation are feasible as well).

To measure completeness, formulas (2a) and (2b) are used. In this example, we assume that the set of attribute values which are considered as incomplete $S_{Incomp}^{a_m}$ is equal to *NULL:* $S_{Incomp}^{a_m} = \{NULL\}$. The metric can hence be implemented without much effort by a SQL Statement of the form "SELECT C_ID FROM customers WHERE e_mail_address IS NULL". Thus, the result of the metric for completeness for the records returned by this statement equals 0 and for the remaining records 1.

To determine the validity of an e-mail-address, formulas (3a) and (3b) are applied. An e-mail address shall be considered valid only if its top-level domain (e.g. .com, .de, .it, etc.) corresponds to a given set of top-level domains. Simplifying, we assume that all records which are returned by the following SQL statement are valid and constitute therefore $S_{Vali}^{a_m}$: "SELECT C_ID FROM customers WHERE e_mail_address LIKE ‚%.org' OR ‚%.com' OR ‚%.aero' OR ‚%.biz' OR ‚%.cat' OR ‚%.com' OR ‚%.coop' OR ‚%.edu' OR ‚%.gov' OR ‚%.info' OR ‚%.int' OR ‚%.jobs' OR ‚%.mil' OR ‚%.mobi' OR ‚%.museum' OR ‚%.name' OR ‚%.net' OR ‚%.org' OR ‚%.pro' OR ‚%.travel". Thus, the result of the metric for validity for the records returned by this statement equals 1. For all other records the result of the metric equals 0.

As mentioned earlier, it is not necessary to define a value range to measure currency. Instead, the distribution function of the shelf-life S^{a_m} has to be determined. The distribution's parameters can be determined in an objective way by statistical methods on the basis of random samples, statistical distributions, and historical data or in a subjective way by expert estimates (a detailed procedure to develop metrics for currency can be found in Heinrich et al. [24]). As soon as a suitable distribution function and parameters have been determined, an automated measurement can be conducted repeatedly. Therefore, the time span which has elapsed since the attribute values' storage in the database, has to be determined first, based on their entry dates. The respective entry dates again can be selected by a SQL statement ("SELECT C_ID, entry_date FROM customers"). The difference between the instant of measuring currency and the e-mail addresses' entry dates results in the time span of

interest $s^{v_I(t_n.a_m)}$. Using this time span and the distribution parameters, the respective currency of an e-mail address can be calculated. Hereafter we exemplarily assume that the shelf life of an e-mail address is exponentially distributed with a decline rate of 0.1. The latter indicates how many values of the attribute a_m become out-of-date on average within one period of time. Thus, we obtain the following metric for currency:

$$Q_{curr}(s^{v^{t_n,a_m}}) := \exp(-0.1 \cdot s^{v^{t_n,a_m}}) \quad (11)$$

Within the fourth and last step the results from these three metrics have to be multiplied for each e-mail address according to formula (7). For the e-mail addresses from Table 1 we obtain the results for the indicator function depicted in Table 2. We also list the results when using the weighted average as aggregation function (cf. section 2.2) to discuss the differences and exclude thereof the dimensions 'accuracy', 'interpretability' and 'accessibility' for the reasons discussed earlier. Moreover, we assume that the remaining dimensions considered are equally weighted ($w_{Comp} = w_{Vali} = w_{Curr} = 1/3$).

Table 2 Results for the indicator function on 2012-03-15

C_ID	e_mail_address	entry_date	Q_{Comp}	Q_{Vali}	Q_{Curr}	Q_{Ind}	Q_{Over}
1	O.Hansen@example.com	1998-11-01	1	1	0.27	0.27	0.76
2	p.parker@world-time.time	2010-01-17	1	0	0.80	0.00	0.60
3	*NULL*	2007-09-27	0	0	0.64	0.00	0.21

The results from Table 2 show that two out of the three e-mail addresses are for sure inaccurate, as they are either not valid (C_ID 2) or incomplete (C_ID 3). Consequently, the value for Q_{Ind} is 0 in both cases. In contrast, Q_{Over} is greater than 0 in both cases; for customer 2, the quality is, in fact, judged quite good at a value of $Q_{Over} = 0.6$. The third e-mail address (customer with C_ID 1) could not be identified as inaccurate according to the indicator function. But, the result indicates that the attribute value is only up-to-date with a probability of 27% which is also the value of the indicator function. The weighted average rates the quality of this attribute value however much higher at 0.76, as the relatively low value for currency is compensated by the high values for completeness and validity. The results of the indicator function can be used as an input to formula (8) in order to calculate the DQ of the attribute "e_mail_address". The resulting value of the indicator function is 0.09. That is, the average quality of the attribute "e_mail_address" is only 0.09.

Transferring this comparatively simple example to very large datasets shows the potential of this automated, repeated, and practicable indication of accuracy. Nonetheless, this is just an example, which demonstrates its applicability and advantages compared to existing approaches. Thus, further attempts are needed to evaluate this indicator function in real-world settings to gain further information from an economic point of view.

5 Conclusion

This paper contributes to an economics-oriented DQ management by designing an indicator function for accuracy as well as by defining eight requirements for aggregating metrics for the DQ dimensions completeness, validity, and currency. The indicator function is based on metrics for the DQ dimensions completeness, validity, and currency, which are aggregated by means of the algebraic product. This procedure enables for a predominantly automated measurement of accuracy. Owing to the avoidance of cost intensive real-world test, this is an advantage especially in very large data sets and in recurring measurements. The indicator function results from a requirements-driven design, ensuring an inter-subjectively verifiable and scientifically founded search process. Besides, formulas to indicate DQ on the levels of attributes, relations, and the database itself are developed. The general applicability of the indicator function is demonstrated in a customer database example and its results are compared to existing approaches.

Some limitations provide room for further research. One limitation exists regarding the dimension 'validity'. As described earlier, the valid value domain of an attribute value $v_l(t_n.a_m)$ can also be derived from another attribute value $v_l(t_n.a_o)$ for attribute a_o of the same tupel t_n. That is, interdependencies among different attribute values of the same tuple can be used to determine an attribute value's validity. However, this procedure has one shortcoming which has not been solved yet: Before determining an attribute value's validity depending on another attribute value of the same tuple, the quality of the latter attribute value has to be determined first (cf. section 2.1.3). Consequently, further research is needed to define metrics for measuring validity taking into account such dependencies. Another limitation of this paper is the missing empirical evidence. Currently, only an example demonstrates the indicator function's general applicability. To further validate the indicator function and its results, several case studies should be conducted. The authors are currently working on an application of the indicator function in the context of managing address data. Results from this study may provide further insides on the costs and benefits of the indicator function under real-world conditions.

References

1. Ballou, D. P., Pazer, H. L.: Modeling completeness versus consistency tradeoffs in information decision contexts. IEEE Trans.Knowled.Data Eng. 1, 240-243 (2003)
2. Ballou, D. P., Pazer, H. L.: Designing information systems to optimize the accuracy-timeliness tradeoff. Information Systems Research. 1, 51-72 (1995)
3. Ballou, D. P., Tayi, G. K.: Enhancing Data Quality in Data Warehouse Environments. Communications of the ACM. 1, 73-78 (1999)
4. Ballou, D. P., Wang, R. Y., Pazer, H. L., Tayi, G. K.: Modeling Information Manufacturing Systems to Determine Information Product Quality. Management Science. 4, 462-484 (1998)

5. Batini, C., Barone, D., Cabitza, F., Grega, S.: A Data Quality Methodology for Heterogenous Data. International Journal of Database Management Systems. 1, 60-79 (2011)
6. Batini, C., Scannapieco, M.: Data Quality. Concepts, Methodologies and Techniques (Data-Centric Systems and Applications) , vol. 1, Berlin (2006)
7. Blake, R., Mangiameli, P.: The Effects and Interactions of Data Quality and Problem Complexity on Classification. Journal of Data and Information Quality (JDIQ). 2, 8 (2011)
8. Calero, C., Caro, A., Piattini, M.: An applicable data quality model for web portal data consumers. World Wide Web. 4, 465-484 (2008)
9. Cappiello, C. and Comuzzi, M. A Utility-Based Model to Define the Optimal Data Quality Level in IT Service Offering. In: Proceedings of the 17th European Conference on Information Systems (ECIS), pp. 1062-1074. Verona (Italy) (2009)
10. Cappiello, C., Francalanci, C., Pernici, B.: Time-Related Factors of Data Quality in Multichannel Information Systems. Journal of Management Information Systems. 3, 71-91 (2004)
11. Caro, A., Calero, C. and Piattini, M. Development process of the operational version of PDQM, LNCS 4831, pp. 436-448. Springer-Verlag, (2007)
12. Codd, E. F.: Extending the database relational model to capture more meaning. ACM Transactions on Database Systems (TODS). 4, 397-434 (1979)
13. CSO Insights: 2005 Executive Report: Target Marketing Priorities Analysis. (2005)
14. De Amicis, F., Barone, D. and Batini, C. An analytical framework to analyze dependencies among data quality dimensions. In: Proceedings of the 11th International Conference on Information Quality (ICIQ), pp. 369-383. Cambridge, MA, (USA) (2006)
15. Eppler, M. J.: Managing information quality, vol. 1, Berlin (2003)
16. Even, A., Shankaranarayanan, G.: Utility-Driven Assessment of Data Quality. The DATA BASE for Advances in Information Systems. 2, 75-93 (2007)
17. Even, A. and Shankaranarayanan, G.: Value-driven data quality assessment. In: Proceedings of the 10th International Conference on Information Quality (ICIQ), pp. 221-236. MIT Press, Cambridge, MA, (USA) (2005)
18. Even, A., Shankaranarayanan, G., Berger, P. D.: Economics-Driven Data Management: An Application to the Design of Tabular Datasets. IEEE Transactions on Knowledge and Data Engineering. 6, 818-831 (2007)
19. Fisher, C. W., Chengalur-Smith, I. N., Ballou, D. P.: The Impact of Experience and Time on the Use of Data Quality Information in Decision Making. Information Systems Research. 2, 170-188 (2003)
20. Fox, C., Levitin, A., Redman, T. C.: The Notion of Data and Its Quality Dimensions. Information Processing & Management. 1, 9-19 (1994)
21. Gackowski, Z. J. Logical interdependence of data/information quality dimensions—A purpose-focused view on IQ. In: Proceedings of the Ninth International Conference on Information Quality (ICIQ 2004) Cambridge, MA, (USA) (2004)
22. Görz, Q.: An Economics-Driven Decision Model for Data Quality Improvement – A Contribution to Data Currency. In: Proceedings of the 17th Americas Conference on Information Systems (AMCIS) Detroit, Michigan (USA) (2011)

23. Information Workers Beware: Your Business Data Can't Be Trusted, http://www.sap.com/about/newsroom/businessobjects/20060625_005028.epx
24. Heinrich, B., Kaiser, M., Klier, M.: A Procedure To Develop Metrics For Currency And Its Application In CRM. ACM Journal of Data and Information Quality. 1, 5:1-5:28 (2009)
25. Heinrich, B., Kaiser, M. and Klier, M.: Does the EU Insurance Mediation Directive help to improve Data Quality? - A metric-based analysis. In: Proceedings of the 16th European Conference on Information Systems (ECIS) Galway, (Irland) (2008)
26. Heinrich, B., Kaiser, M. and Klier, M.: How to measure data quality? – a metric based approach. In: Proceedings of the 28th International Conference on Information Systems (ICIS) Montreal, (Canada) (2007)
27. Heinrich, B., Kaiser, M. and Klier, M.: Metrics for measuring data quality – Foundations for an economic data quality management. In: 2nd International Conference on Software and Data Technologies (ICSOFT) Barcelona, (Spain) (2007)
28. Helfert, M., Foley, O., Ge, M. and Cappiello, C.: Limitations of Weighted Sum Measures for Information Quality. In: Proceedings of the 15th Americas Conference on Information Systems (AMCIS) San Francisco, CA, (USA) (2009)
29. Juran, J. M.: How to think about Quality. , vol. 5, 2.1-2.18, New York (1998)
30. Kahn, B. K., Strong, D. M., Wang, R. Y.: Information quality benchmarks: product and service performance. Commun ACM. 4, 184-192 (2002)
31. Lee, Y. W., Pipino, L., Strong, D. M., Wang, R. Y.: Process-Embedded Data Integrity. Journal of Database Management. 1, 87-103 (2004)
32. Lee, Y. W., Strong, D. M., Kahn, B. K., Wang, R. Y.: AIMQ: a methodology for information quality assessment. Information & Management. 2, 133-146 (2002)
33. Naumann, F., Freytag, J., Leser, U.: Completeness of Integrated Information Sources. Information Systems. 7, 583-615 (2004)
34. Orr, K.: Data Quality and Systems Theory. Communications of the ACM. 2, 66-71 (1998)
35. Otto, B., Lee, Y. W., Caballero, I.: Information and data quality in business networking: a key concept for enterprises in its early stages of development. Electronic Markets. 83-97 (2011)
36. Parssian, A., Sarkar, S., Jacob, V. S.: Assessing Data Quality for Information Products: Impact of Selection, Projection, and Cartesian Product. Management Science. 7, 967-982 (2004)
37. Pipino, L., Lee, Y. W., Wang, R. Y.: Data Quality Assessment. Communications of the ACM. 4, 211-218 (2002)
38. Russom, P.: Taking Data Quality to the Enterprise through Data Governance. The Data Warehousing Institute, Seattle. Seattle (2006)
39. Vassiliou, Y.: Null values in data base management - a denotational semantics approach. In: Proceedings of the 1979 ACM SIGMOD International Conference on Management of Data (SIGMOD '79), pp. 162-169. ACM, Boston, (USA) (1979)
40. Wand, Y., Wang, R. Y.: Anchoring data quality dimensions in ontological foundations. Communications of the ACM. 11, 86-95 (1996)

41. Wang, R. Y.: A Product Perspective on Total Data Quality Management. Communications of the ACM. 2, 58-65 (1998)
42. Wang, R. Y., Strong, D. M.: Beyond accuracy: what data quality means to data consumers. Journal of Management Information Systems. 4, 5-33 (1996)