A Measure-Theoretic Foundation for Data Quality

Antoon Bronselaer, Robin De Mol, and Guy De Tré

***Abstract*—In this paper, a novel framework for data quality measurement is proposed by adopting a measure-theoretic treatment of the problem. Instead of considering a specific setting in which quality must be assessed, our approach departs more formally from the concept of measurement. The basic assumption of the framework is that the highest possible quality can be described by means of a set of predicates. Quality of data is then measured by evaluating those predicates and by combining their evaluations. This combination is based on a capacity function (i.e., a fuzzy measure) that models for each combination of predicates the capacity with respect to the quality of the data. It is shown that expression of quality on an ordinal scale entails a high degree of interpretation and a compact representation of the measurement function. Within this purely ordinal framework for measurement, it is shown that reasoning about quality beyond the ordinal level naturally originates from the uncertainty about predicate evaluation. It is discussed how the proposed framework is positioned with respect to other approaches with particular attention to aggregation of measurements. The practical usability of the framework is discussed for several well known dimensions of data quality and demonstrated in a use-case study about clinical trials.**

***Index Terms*—Data quality, fuzzy measure, uncertainty modeling.**

I. Introduction

THE continuously growing potential of data in nowadays organizations has rapidly promoted assessment of data quality to an important topic of research. Throughout the past decades, many authors have contributed to this field and some commonly accepted principles have been established. One of these principles is that data quality is a *multidimensional* problem [1], [2]. Although many such dimensions may be relevant in particular situations [3], [4], the most commonly studied dimensions are correctness, completeness, and consistency, and time-related dimensions such as timeliness, currency, and volatility [5], [6]. For each of these dimensions, many procedures for measurement have been proposed with very specific scenarios.[1] Because the different dimensions of quality typically have their particular nature, the whole of data quality measures across all dimensions is very disperse and heterogeneous. As a consequence, a common and formal understanding of the concept "measurement" in the field of data quality is up-to-date still missing, as are any well-established connections to the theory of representational measurement [7]. To the best of our knowledge, a first observation in this direction was made by Even *et al.* [8]–[10], who compiled a set of properties (i.e., axioms) to which a "good" measure of quality should adhere. Their perspective was mainly economically and utility driven. The considered properties were further investigated and refined by Heinrich *et al.* [11] from a decision making point of view. Although these properties are valuable and interesting, the requirements in [8] and [11] originate from a specific usage scenario. Therefore, it is argued here that a more general approach is required. Recently, observations of the same issue were made in [12] and [13]. The need for a more general approach can be shown by reviewing some problems with state-of-the-art measurement approaches.

First, definitions of measurement are very heterogeneous: some are based on metrics [14], some are based on utility calculation [8], and others apply a function on the data [15]. It is therefore hard to compare different approaches as they express quality in a different way. In this respect, Fürber *et al.* [16] proposed an ontological approach in the definition of data quality rules. However, the authors believe that the field of data quality measurement has not yet reached sufficient maturity for such an approach.

Second, the commonly accepted idea of expressing quality in the unit interval $[0, 1]$ (or an alternative isomorphic scale) causes issues with the interpretation of quality measurements. What does it mean when an attribute value has an accuracy of 0.7 or when the timeliness is 0.8? Moreover, if we assume to operate on an interval scale, then what is the *unit* of measurement? In general, numbers assigned to indicate quality are hard to interpret and they do not provide insight in the *causality* of quality degradation, nor do they enable concrete actions that must be taken to *improve* the quality of current or future data. Again, the lack of a theoretical framework for measurement is at the root of this problem. There are no procedures that tell us how to properly assign numbers to data in such a way it reflects our empirical perception of quality. If it is unclear what measuring means, the actual measurements cannot easily be interpreted, nor can they be combined.

Third, existing approaches assume that quality is measured on the level of attributes and then aggregated to higher levels. In order to illustrate that such an approach can easily fail, consider the snippet of data shown in Table I that shows a number of addresses in the city of Ghent. Hereby, blank cells indicate a NULL value. Note that, for each of the addresses shown in Table I, there is at most one attribute that has a NULL value.

Manuscript received May 27, 2016; revised October 17, 2016; accepted January 18, 2017. Date of publication March 23, 2017; date of current version March 29, 2018. *(Corresponding author: Antoon Bronselaer.)*

The authors are with the Department of Telecommunications and Information Processing, Ghent University, Ghent B-9000, Belgium (e-mail: antoon.bronselaer@ugent.be; robin.demol@ugent.be; guy.detre@ugent.be).

Color versions of one or more of the figures in this paper are available online at http://ieeexplore.ieee.org.

Digital Object Identifier 10.1109/TFUZZ.2017.2686807

[1] In the literature, the term "metric" is sometimes used. Here, the term "measure" is preferred as a metric is commonly used to express a notion of distance.

TABLE I
EXAMPLE DATA FOR QUALITY MEASUREMENT

ID	Street	Number	City	Zip
1	Saint-Bavo Square	25	Gent	9000
2	Saint-Bavo Square	40		9000
3	Saint-Bavo Square	41	Gent	
4	Saint-Bavo Square	41	Gent	2000
5	Saint-Bavo Square		Gentbrugge	9000
6	Saint-Bavo Square		Gent	9000

Suppose now we want to measure completeness. Following the approach proposed in [17], completeness is measured on the attribute level by means of the function:

$$c(x) = \begin{cases} 1, & \textbf{if} \quad x \neq \text{NULL} \\ 0, & \textbf{if} \quad x = \text{NULL} \end{cases}.$$

It is further stated in [17] that aggregation to the tuple level is then done by taking the average c over all attributes. Doing this yields a score of 1 for tuples 1 and 4 and a score of 0.8 for all other tuples. Following the approach in [18], a tuple is scored 0 if it contains at least one NULL value. With this approach, a similar result is obtained in the sense that tuples 2, 3, 5, and 6 are assigned with the same level of quality. However, these approaches omit that certain dependencies between data can exist. For Belgian postal addresses, there is a kind of redundancy between the zip code and the name of the city. Therefore, provided that the zip code is known, a NULL value for street or number reduces the quality of the data *more* than a NULL value for the name of the city.

A fourth and last issue deals with aggregation of different measurements on the same data. In order to assess the usefulness of data with respect to a certain goal, it is often required to measure multiple dimensions of quality. As the data in Table I concern address data, a concrete goal could be to assess whether an address is of sufficient quality so that postal mail can be directed to this address. From this point of view, tuple 4 has perfect completeness, but it is useless because there is an inconsistency between the name of the city and the zip code. This example illustrates the need for measuring multiple aspects of quality at the same time in order to decide about the usability of data. However, current approaches for measurement of data quality do not account for the fact that different measurements may and usually will have a different nature and interpretation. The fact that all measurements are expressed in the unit interval often cloaks this difference in interpretation. Within the framework proposed here, this problem will be given special attention.

To cope with the above mentioned problems, a measure-theoretic treatment of data quality is proposed in this paper. Instead of starting of, by considering specific dimensions or applications, the concept of measurement is cut loose here from any context. We propose a framework for measurement that is applicable to specific scenarios. In contrast to the axiomatic definition of quality measures in [8], [11], a more general approach based on *predicates* and *capacity* is envisioned here. Within this framework, it is argued that capacity should be expressed using a scale that is *ordinal* [19]. Scales beyond the ordinal level are not considered here because they unavoidably introduce the problem of choosing a "unit" of quality. By conducting a purely measure-theoretic inference, a framework for quality measurement is obtained in which a measurement is unambiguous and therefore highly informative. Several results with respect to interpretation and representation will be shown and a comparison with the axiomatic ways of defining quality measurement will be made. In this comparative discussion, particular attention will be given to the concept of aggregation.

The remainder of this paper is structured as follows. In Section II, the literature relevant to the current paper is reviewed and the proposed framework is positioned with respect to these contributions. In Section III, some notations for relevant preliminary concepts such as the relational database model are introduced. In Section IV, a measure-theoretic framework for the measurement of data quality is introduced and the properties of this framework are investigated. A theoretical comparison with other approaches is conducted and it is explained how quantitative information about data quality arises when uncertainty about predicates is added to the picture. In Section V, the proposed procedure for measurement is applied to different well known dimensions. In Section VI, a use-case for consistency about clinical trials is reported. Finally, the most important contributions of this paper are summarized in Section VII.

II. RELATED WORK

In the past decades, several authors have contributed to the measurement, assessment, and improvement of data quality. In this section, the most relevant contributions with respect to the current paper are summarized.

The first contributions toward the multidimensional quality model that is today's standard, are due to Wang *et al.* [1], [20] and Redman [5]. In their work, they argue for the need of well-defined, goal-oriented dimensions of data quality. This multidimensional view inspired several authors to define a broad range of different data quality dimensions. Kim *et al.* proposed a taxonomy of different quality dimensions [4] and Batini *et al.* investigated the more common dimensions and how to measure them [2], [6]. When it comes to defining *measures* for data quality dimensions, Pipino *et al.* argued that a distinction can be made between *objective* and *subjective* measures [3]. This distinction was further developed by Even *et al.* [8]–[10] who point out the distinction between impartial, context-free measures on the one hand and contextual, utility-driven measures on the other hand. Of particular importance in the work by Even *et al.* is their proposal of a set of requirements to which measures for data quality should adhere. This set of requirements was adopted and refined by Heinrich *et al.* [14], who provide an axiomatic definition of a data quality measure by stating six axioms:

1) *Normalization:* A measure must be adequately normalized and expressed using a *bounded* scale.
2) *Interval Scaled:* A measure must be expressed in an interval-scale to support both monitoring (e.g., over time) and economic assessment of the measure.

3) *Interpretability:* A measure must be comprehensible and "easy to interpret by business users" [10].
4) *Adaptivity:* As data quality is often measured in a specific context, measures should be adaptable to such context (e.g., by means of parameters).
5) *Feasibility:* A measure must be based on input parameters that are determinable and allow for a high level of automation.
6) *Aggregation:* A measure should allow aggregation along different structural levels of a (relational) database (i.e., attribute level, tuple level, relation level, and database level).

In their motivation for these axioms, Heinrich *et al.* adopt an economically driven context of decision making. In the remainder of this paper, we will further discuss these axioms when reviewing our framework. Recently, several authors have attempted to standardize data quality terminology by means of ontologies [16], [21]. In the spirit of this paper, measurement of data quality as proposed here is aimed to be standardized, portable, and easily exchangeable. Next to the high level and comprehensive contributions reviewed above, there are many contributions that focus on specific dimensions. For example, Heinrich *et al.* have investigated completeness [17], accuracy [17], and currency [11], [22] separately. Naumann *et al.* have investigated completeness in the setting of distributed querying [15]. They assess the usefulness of one or more sources based on their completeness. This idea to incorporate quality measurements in the resolution of a query has recently been investigated in more depth [23]. Ballou *et al.* have studied the tradeoff between concurrent quality dimensions in a decision making context [24], [25]. They argue that, in an economically driven scenario, a lack of perfect data requires a mechanism that can select the best possible data by making a tradeoff between several, possibly conflicting, quality dimensions of the data. They studied the balancing between accuracy and timeliness [24] and the balancing between consistency and completeness [25], [26].

The impact of data quality on other areas of research has been studied by several authors. Ballou *et al.* investigated data quality in the scope of data warehouses [27]. Blake *et al.* studied the most important dimensions from the perspective of data mining [18]. Caballero *et al.* reviewed data quality in the setting of a big data scenario and pose that the most important dimension in big data projects is consistency [28]. With respect to the big data scenario, Abedjan *et al.* [29] pointed out that the big data scenario poses concrete challenges for data quality research. Among others, they indicate that incremental algorithms for measurement are an important topic of future research. When discussing aggregation of measurements, attention will be given to this point. To conclude this literature review, we mention that some authors surpass the process of measurement and aim at data quality improvement. Chen *et al.* [30] have investigated the repairing of functional dependency violations to improve consistency. Cong *et al.* investigated a similar principle based on conditional functional dependencies [31].

III. Notations

In the remainder of this paper, the relational database model is considered [32]. Note that although we consider the relational model here, the presented concepts can be transferred easily to nonrelational database systems such as NoSQL and XML databases. In this section, the relevant concepts and notations used throughout the paper are recalled. Consider a countable set of attributes $\mathcal{A}$, where each attribute $a \in \mathcal{A}$ is defined by a *name* and a *domain*. For each attribute $a \in \mathcal{A}$, $\mathrm{dom}(a)$ denotes the domain of a and $\mathrm{name}(a)$ denotes the name of a. In practice, if no ambiguity exists, it is common to omit the name function and use the notation a for both the name of the attribute as well as the attribute itself. Given a set of attributes, a *(relational) schema* $\mathcal{R}$ is defined by a *nonempty* and *finite* subset of $\mathcal{A}$. Instantiations of a schema are known as *relations*, where a relation R over $\mathcal{R}$ is defined by $R \subseteq \mathrm{dom}(a_1) \times \cdots \times \mathrm{dom}(a_k)$. In other words, a relation over a schema can be any subset of the crossproduct of all attribute domains in the schema. Each element of a relation R with schema $\mathcal{R}$ is called a tuple t over $\mathcal{R}$. Such a tuple t is basically a vector where the ith dimension contains a value for attribute a_i. The relational model comes with a complete relational algebra, but within the scope of this paper only the projection operator is of relevance. For a relation R with schema $\mathcal{R}$, the projection of R over a set of attributes $A \subseteq \mathcal{R}$ is denoted as $R[A]$ and is defined as a relation with schema A that consists of the set of tuples from R that is obtained by projecting R over attributes in A. In the case where R and A are given by the singleton sets $\{t\}$ and $\{a\}$, the notation $t[a]$ is adopted.

IV. Measure-Theoretic Approach to Quality Measurement

Having the scope, the goals and the notations set, a measure theoretic framework for quality is now proposed in this section. The properties of this framework are investigated and their practical relevance is discussed. The section then continues with discussing some ways to facilitate the construction of capacities. Finally, a comparison with the axiomatic definition of data quality measurement is presented.

A. Basic Framework

In order to provide a formal ground for measuring data quality, we begin by specifying the scope of data quality measurement. In the literature, many approaches adopt the assumption that data quality should be measured on the lowest level possible and then be aggregated to higher levels [10], [15], [17]. However, it is argued here that, in a more general setting, it makes sense to immediately consider higher level structures. The example of measuring address completeness (see Section I) illustrates that considering multiple attributes at once might be necessary. In addition, when verifying uniqueness of attributes or (conditional) functional dependencies, it might be necessary to consider multiple tuples at once. Therefore, in general, the data that must be assessed are given here by a relation R with schema $\mathcal{R}$. Note that the scope in which quality is measured, be it on attribute level, tuple level or relational level, is something

that should be considered upfront and will have consequences, e.g., when it comes to computing aggregates of quality measurements.

In order to measure the quality of R, the basic assumption on which our framework relies is that R is of the best possible quality if it satisfies a finite set of criteria. These criteria must be at least *sufficient* in the sense that satisfaction of criteria in addition to the given ones, does not improve the quality of R any further. Within the scope of this paper, it is assumed that each criterion is represented by means of a predicate $p : \mathrm{dom}(\mathcal{R}) \rightarrow \mathbb{B}$ where $\mathbb{B} = \{T, F\}$ is the set of Boolean values on which we consider the usual ordering $F < T$. Given a finite set of requirements, the corresponding predicates are denoted by $P = \{p_1, \ldots, p_n\}$. The choice for Boolean-valued criteria is motivated by two important arguments. First, as will be shown in the following, measurement of predicates in the quality measurement space can be done without any ambiguity. This yields an exact interpretation of quality measurement. Second, on a more intuitive level, data quality has a strong affinity with the Boolean space. In the end, measurement of quality will serve to answer the question "Are the data of sufficient quality?" From this perspective, a Boolean treatment at the basis is justified by intuition.

Basically, *measurement* of the quality of R is an appreciation of predicates observed on R. This appreciation will be expressed on an ordinal scale $\mathbb{S}$ that has at least two levels. We do not consider interval scales (or higher) as this would imply the choice of a data quality "unit" and this is a problem of which it is yet unclear whether it can be solved. The total order of $\mathbb{S}$ is denoted as $\leq$. The smallest (resp. largest) element of $\mathbb{S}$ under $\leq$ is denoted as $\mathbb{0}$ (resp. $\mathbb{1}$). Measurement of quality is then defined by a function $Q : \mathrm{dom}(\mathcal{R}) \rightarrow \mathbb{S}$. Because of the assumption that P represents a set of sufficient requirements for data of the best possible quality, we must have

$$\left(\bigwedge_{p \in P} p(R) \right) \Rightarrow Q(R) = \mathbb{1}. \tag{1}$$

If it would be assumed that in addition, P models a set of *necessary* requirements, then we must have

$$\left(\bigwedge_{p \in P} p(R) \right) \Leftrightarrow Q(R) = \mathbb{1}. \tag{2}$$

In order to construct Q, we will operate within the measurable space $\left(P, 2^P\right)$ and transfer the concept of *capacity* [33] to the context of data quality. A definition for quality capacity is therefore introduced as follows.

Definition 1 (Quality Capacity): Consider the data R and let P be a set of predicates that are sufficient requirements for data of the best possible quality. Consider an ordinal scale $\mathbb{S}$ equipped with a total order $\leq$. A quality capacity on P is defined by a function $\mathcal{C} : 2^P \rightarrow \mathbb{S}$ for which $\mathcal{C}(\emptyset) = \mathbb{0}, \mathcal{C}(P) = \mathbb{1}$ and that is monotonic in the sense that

$$\forall X_1 \subseteq P : \forall X_2 \subseteq P : X_1 \subseteq X_2 \Rightarrow \mathcal{C}(X_1) \leq \mathcal{C}(X_2). \tag{3}$$

Informally, a quality capacity (also known as a *fuzzy* measure [34]) expresses for any combination of predicates, the perceived quality when at least these predicates succeed. In other words, for a subset of predicates $X \subseteq P, \mathcal{C}(X)$ represents the maximal capacity of these predicates with respect to the quality of data. Because a capacity has no notion of *additivity*, the codomain is allowed to be ordinal scaled.

Before we get to the integral-based inference of a measurement function Q, some interesting notes on the implications of Definition 1 on P are given. First, a quality capacity is monotonic. This means that observing more predicates cannot decrease the appreciation that we assign to the given data. Second, if we require P to be necessary and sufficient requirements, then $\mathcal{C}$ must satisfy $\mathcal{C}(X) = \mathbb{1} \Leftrightarrow X = P$. A quality capacity that satisfies this constraint is called a *strict* quality capacity.

Let us now consider the measurable space $\left(P, 2^P\right)$. In order to measure quality, it is required to formalize how predicate evaluation contributes to quality and how it is to be measured in $\mathbb{S}$. This formalization is obtained by means of a $\mathcal{C}$-measurable function $h : P \rightarrow \mathbb{S}$ defined by:

$$h(p) = \begin{cases} \mathbb{1}, & \textbf{if} \quad p \models T \\ \mathbb{0}, & \textbf{if} \quad p \models F \end{cases}. \tag{4}$$

This function is basically a characteristic function and states that if a predicate evaluates to T, this is measured as perfect in $\mathbb{S}$. Despite its deceptively simple formulation, the definition of h deserves special attention. The choice for predicates makes the measurement of predicates in $\mathbb{S}$ completely free of ambiguity. If criteria would be evaluated by functions with a codomain that is larger than $\mathbb{B}$ (e.g., $[0, 1]$), the meaning of the evaluation of a predicate could easily differ across different predicates and the construction of h would be much more ambiguous. As such, our earlier remark about the interpretation of predicates is supported here on a more formal level. Given a piece of data, the quality of that data can now be calculated by integrating the function h over the set of predicates P with respect to the capacity $\mathcal{C}$:

$$Q(R) = \int_P h(p) \circ \mathcal{C}. \tag{5}$$

The calculation of integrals in an ordinal setting was investigated by Sugeno in his doctoral dissertation [35]. Following the inference rules of the Sugeno integral, it can be shown that

$$Q(R) = \sup_{\alpha \in \mathbb{S}} \left(\min\left(\alpha, \mathcal{C}(S_\alpha)\right) \right) \tag{6}$$

where we have that $S_\alpha = \{p \mid p \in P \wedge h(p) \geq \alpha\}$. Informally, this integral looks for those predicates that have maximal *measured* capacity (under h) and at the same time have maximal potential capacity under $\mathcal{C}$. Because of the choice for predicates, the definition of h allows us to further simplify the calculation of Q. Indeed, because h is a function with a binary image $\{\mathbb{0}, \mathbb{1}\}$, S_α is restricted to two possible values. More specifically, if $\alpha = \mathbb{0}$, then $S_\alpha = P$. Alternatively, if $\alpha \neq \mathbb{0}$, then we have $S_\alpha = \{p \mid p \in P \wedge p \models T\}$. This implies that

$$Q(R) = \sup\left(\mathbb{0}, \mathcal{C}\left(\{p \mid p \in P \wedge p(R) = T\}\right)\right). \tag{7}$$

As such, the measured quality of data for R is given by

$$Q(R) = \mathcal{C}\left(\{p \mid p \in P \wedge p(R) = T\}\right). \tag{8}$$

It can be seen that this measure-theoretic inference results in an elegant and simple formulation of quality measurement.

B. Properties

In the following, the properties of Q with respect to the interpretation of measurements are further investigated and reported. Let us begin by noting that if P is finite then the number of possible outcomes for Q is also finite and bounded by the number $2^{|P|}$. In combination with this upper bound, the following property offers a great potential with respect to interpretation.

Property 1: Consider a measurement function Q based on predicates P and a capacity $\mathcal{C}$ that maps onto $\mathbb{S}$. If $Q(R) = s$ with $s \in \mathbb{S}$, then, we have

$$\forall p \in \left(\bigcap_{X \subseteq P \wedge \mathcal{C}(X) = s} X \right) : p(R) = T \tag{9}$$

and

$$\forall p \in \left(P \setminus \bigcup_{X \subseteq P \wedge \mathcal{C}(X) = s} X \right) : p(R) = F. \tag{10}$$

Proof: Assume that $Q(R) = s$. We, then, have

$$\exists X \subseteq P : \mathcal{C}(X) = s \wedge (\forall p \in X : p(R) = T). \tag{11}$$

On the one hand, if there exists a predicate p' such that

$$\forall X \subseteq P : \mathcal{C}(X) = s \Rightarrow p' \in X \tag{12}$$

then, we are certain that p' evaluates to true for the given data. On the other hand, if there exists a predicate p'' such that

$$\forall X \subseteq P : \mathcal{C}(X) = s \Rightarrow p'' \notin X \tag{13}$$

then, there does not exists an evaluation of predicates under which p'' evaluates to true for the given data. Hence, p'' must evaluate to false. ■

Together with the upper bound on the number of outcomes for Q, Property 1 illustrates that each of the outcomes immediately carries an interpretation. Within the context of the capacity $\mathcal{C}$, observation of $Q(R)$ immediately allows us to draw conclusions about the data. The extent of these conclusions depends on the structure of the capacity $\mathcal{C}$: the more predicate sets map to the same $s \in \mathbb{S}$, the weaker the conclusions about R will be. However, if there is only one subset of predicates that maps to a given s, then observation of s immediately reveals the truth values of *all* the predicates. Therefore, as a corollary we have that, if $\mathcal{C}$ is an injection, then observation of $Q(R)$ encodes the truth values of all predicates for R. Property 1 learns that each value $s \in \mathbb{S}$ comes with certain constraints that need to be satisfied in order to reach the level of quality represented by s. This observation hints us that it should be possible to represent the constraints for each $s \in \mathbb{S}$ as a Boolean function. In order to get to this representation theorem, some intermediary concepts are required.

First, we say that a set of predicates $B \subseteq P$ is *minimal* and *sufficient* for $s \in \mathbb{S}$ if the capacity of B is at least s and no real subset of B has a capacity that is at least s. As such, for each $s \in \mathbb{S}$ the set of all minimal and sufficient predicate sets can be written as

$$\mathcal{B}(s) = \{B \mid B \subseteq P \wedge \mathcal{C}(B) \geq s \wedge \forall B' \subset B : \mathcal{C}(B') < s\}.$$

The minimal and sufficient generator for $s \in \mathbb{S}$ is then defined by

$$G(R \mid s) = \bigvee_{B \in \mathcal{B}(s)} \bigwedge_{p \in B} p(R). \tag{14}$$

The function G is a Boolean representation of the constraints necessary for R to be at least of quality level s. This claim is formalized and proven by considering the following representation theorem.

Theorem 1 (Representation Theorem): For a measurement function Q based on predicates P and a capacity $\mathcal{C}$ we have

$$Q(R) = \max\{s \mid s \in \mathbb{S} \wedge G(R \mid s) = T\}. \tag{15}$$

Proof: Consider the data R and assume that $Q(R) = s$, then by definition there must exist $X \subseteq P$ for which

$$\mathcal{C}(X) = s \wedge \forall p \in X : p(R) = T \tag{16}$$

By definition of $\mathcal{B}(s)$ and by monotonicity of $\mathcal{C}$, we have

$$\exists B \in \mathcal{B}(s) : B \subseteq X \tag{17}$$

from which it follows that

$$\forall p \in B : p(R) = T. \tag{18}$$

By definition, it follows that $G(R \mid s) = T$. Consider now an $s' \in \mathbb{S}$ such that $s' > s$ and assume that $G(R \mid s') = T$. This would imply that there is a $B' \subseteq P$ for which

$$\mathcal{C}(B') \geq s' > s \tag{19}$$

and

$$\forall p \in B' : p(R) = T. \tag{20}$$

By monotonicity of $\mathcal{C}$ we would then have that $Q(R) \neq s$, which is in contradiction with the premises. As such, we have proven that if $Q(R) = s$, then s is the largest value in $\mathbb{S}$ for which $G(R \mid s) = T$. ■

Theorem 1 is an important and elegant result with respect to quality measurement. It shows how measurement of quality verifies for each level of quality, whether the criteria for that level are satisfied, or not. The largest level of quality for which all criteria are satisfied determines the quality of the data. The representation theorem has some interesting consequences that further support the correctness and intuitiveness of our approach. First, it can be seen that the minimal and sufficient generator for $\mathbb{0}$ is a tautology:

$$\forall R : G(R \mid \mathbb{0}) = T. \tag{21}$$

This is because the capacity of an empty predicate set equals $\mathbb{0}$. As such, data are always at least of quality $\mathbb{0}$. This positions $\mathbb{0}$ as the absolute minimal level of quality. Second, if $\mathbb{S} = \mathbb{B}$ then quality measurement reduces to a regular Boolean function as there are only $\mathbb{0}$ (i.e., F) and $\mathbb{1}$ (i.e., T) to be considered. In addition, if the underlying capacity is strict, then this Boolean function is a conjunction of all the predicates in P.

C. Construction of Capacity

So far a formal derivation of quality measurement has been presented and some properties have been shown. In order to get the framework to work, two steps must be taken: formulation of the predicates and construction of the capacity. Because the image of a capacity has an exponential size in terms of the number of predicates, it can be considered a tedious task. For that reason, some notes on the construction of capacities and possible simplifications of this process are discussed.

To begin with, it is noted that an important factor in the construction of capacity is the extent to which predicates in P are mutually independent. If P contains two or more predicates that are mutually dependent, then this dependency unavoidably influences the capacity. A first kind of dependency that may occur is a logical contradiction. If P contains both p and $\neg p$, then any subset containing both p and $\neg p$ reflects an impossible situation. Usually, the occurrence of such logical contradictions within P indicates one or more disjunctive scenarios in which data are of the best quality. For example, consider two attributes a and b that are consistent if their values are either both even or both odd. In such a case, four predicates are to be considered: "$p_1 = t[a]$ is even," "$p_2 = t[b]$ is even," "$p_3 = t[a]$ is odd," and "$p_4 = t[b]$ is odd" where we have that $p_1 = \neg p_3$ and $p_2 = \neg p_4$. When constructing capacity for these predicates, we want to assess the disjunctive scenarios where either both attribute values are even or both attribute values are odd. In other words, we must pay attention in assignment of capacity to sets $\{p_1, p_2\}$ and $\{p_3, p_4\}$. Supersets of $\{p_1, p_2\}$ and $\{p_3, p_4\}$ correspond to contradictory situations that will never occur and the actual assignment of capacity to those supersets does not matter, as long as it respects monotonicity. From this point of view, the construction of capacity can be simplified in the sense that P can be partitioned into subsets in which no mutual contradictions occur and capacity must be assigned only for subsets of those partitions. A second kind of dependency that will often occur is a logical implication. If P contains two predicates for which $p_1 \Rightarrow p_2$, this dependency again influences the construction of capacity in the sense that $\{p_1\}$ and $\{p_1, p_2\}$ can be assigned the same capacity as they are equivalent. Such logical implications are for example quite common when treating uncertainty about predicates (see Section IV-E). Within the scope of this paper, we will not further detail on the impact of predicate dependencies, but it is important to realize that one should account for them during the design of Q.

If all predicates are mutually independent or one is able to account for existing dependencies in P, there are a number of strategies to soothe the construction of capacity. First, the literature on fuzzy measures and integrals describes several options such as *symmetric* measures [36], λ-measures [35], and possibility/necessity measures [34]. For a further reading on nonadditive measures, the reader is referred to [37]. An interesting way to construct the capacity relies on the fact that in many practical situations, there exists an intuitive order in which predicates should be evaluated. As an example, consider the determination of the normal form of a relational database [32], [38]. Each normal form is hereby paired with a set of predicates, but it only makes sense to verify those predicates if all predicates of all lower normal forms are known to evaluate to true. If such a scenario is encountered, the construction of $\mathcal{C}$ can be reduced to a linear problem. Consider therefore the set of predicates P and consider a total order $\prec$ on P such that $p_i \prec p_j$ means that p_i should be evaluated before p_j. With this order at hand, let us define for each $i \in \{0, \ldots, |P|\}$, $E_{(i)}$ as the subset of P that contains the first i predicates in the order $\prec$. We then have that $E_{(0)} = \emptyset$ and $E_{(|P|)} = P$. These sets allow for the definition of the following class of capacities.

Definition 2 ($\prec$-sensitive quality capacity): Consider the predicates P with evaluation order $\prec$. The quality capacity $\mathcal{C}$ is $\prec$-sensitive if

$$\forall X \subseteq P : \mathcal{C}(X) = \max_{E_{(i)} \subseteq X} \mathcal{C}\left(E_{(i)}\right). \tag{22}$$

Definition 2 states that the capacity of a set of predicates cannot be greater than the capacity of the largest $E_{(i)}$ contained by those predicates. This means that the construction of the capacity is equivalent to the construction of a monotonically increasing function that maps each $E_{(i)}$ onto $\mathbb{S}$, taking into account the boundary constraints of capacities. For each $E_{(i)}$, the capacity is given by that function. For any other predicate set X, $\mathcal{C}(X)$ is given by the capacity of the largest $E_{(i)}$ that is contained by X.

Property 2: If $\mathcal{C}$ is $\prec$-sensitive, it is a necessity measure and therefore minitive. Formally, for any $X \subseteq P$ and $Y \subseteq P$ we have

$$\mathcal{C}(X \cap Y) = \min\left(\mathcal{C}(X), \mathcal{C}(Y)\right). \tag{23}$$

Proof: Follows from Definition 2. ■

The class of $\prec$-sensitive quality capacities has some appealing properties. First, because there are only $|P| + 1$ different $E_{(i)}$, the construction of the capacity is linear in terms of $|P|$. Second, if the generative function that assigns capacity to the $E_{(i)}$-sets is a bijection, then the inverse function is defined and a bijection as well. Therefore, the measurement of quality can be translated into those predicates that succeeded and those predicates that failed. Third, application of Theorem 1 allows to transform *any* capacity into an equivalent $\prec$-sensitive capacity by introducing a new set of predicates such that, for each $s \in \mathbb{S}$, the predicate:

$$p_s(R) = G(R|s) \tag{24}$$

and by considering the evaluation order as follows:

$$\forall s_1 \in \mathbb{S} : \forall s_2 \in \mathbb{S} : s_1 < s_2 \Rightarrow p_{s_1} \prec p_{s_2}. \tag{25}$$

This last result indicates the importance of $\prec$-sensitive capacities as it aids in the standardization and simplification of capacity construction.

As another option to construct capacity, it is recalled that the capacity can be used to account for dependencies between attributes (see Section I). In this case, capacity is constructed by translating the existing dependencies into a capacity function. This is illustrated by means of the following example.

Example 1: In this example, a quality measurement function for completeness is constructed that accounts for redundancy between attributes. The data R considered here reduce to a single

tuple t consisting of a set of attributes for which completeness should be measured as a whole. In the sample data of Table I (see Section I), these attributes are all attributes that constitute an address, i.e., street name, house number, zip code, and city name. For each of the attributes in the data, a single predicate is defined that asserts the attribute value being not a NULL value. More specifically, for each attribute $a \in \mathcal{R}$, we consider a predicate p_a defined by

$$p_a(t) = \begin{cases} T, & \textbf{if} \quad t[a] \neq \text{NULL} \\ F, & \textbf{if} \quad t[a] = \text{NULL} \end{cases}. \tag{26}$$

These predicates together describe the requirements met by data that are perfectly complete. In order to measure the completeness of data, the capacity of any set of predicates must be expressed. In this concrete scenario, the capacity of a set of attributes expresses the completeness of the data in case those attributes are not NULL. To construct this capacity, we consider the set of quality levels $\mathbb{S} = \{\text{BAD}, \text{SUFFICIENT}, \text{PERFECT}\}$ and a total order relation such that BAD $<$ SUFFICIENT $<$ PERFECT. As explained in Section I, the data in Table I contain a certain redundancy between the zip code and the name of the city. This redundancy implies that data for which *only* one of these attributes is missing, has sufficient completeness. In terms for the capacity $\mathcal{C}$, this means that:

$$\mathcal{C}\{p_{\text{street}}, p_{\text{number}}, p_{\text{city}}\} = \text{SUFFICIENT} \tag{27}$$

$$\mathcal{C}\{p_{\text{street}}, p_{\text{number}}, p_{\text{zip}}\} = \text{SUFFICIENT}. \tag{28}$$

Because of the upper boundary constraint of a capacity, we also have

$$\mathcal{C}\{p_{\text{street}}, p_{\text{number}}, p_{\text{city}}, p_{\text{zip}}\} = \text{PERFECT}. \tag{29}$$

For all other attributes sets, the capacity is equal to BAD. If these predicates and capacity function are used to measure the completeness of the tuples in Table I, we find that tuples 1 and 4 have perfect completeness because they have no NULL values among their attributes. Tuples 2 and 3 have sufficient quality, because those tuples have a NULL value for resp. the city name and the zip code. Finally, tuples 5 and 6 have a completeness score equal to BAD because the missing house number makes them useless according to the given capacity. It can be observed that the resulting measurement of completeness is more fine grained than the procedures described in Section I because those procedures yielded an equal completeness measurement for tuples 2, 3, 5, and 6. The above sketched procedure for the measurement of completeness illustrates the borderline between objective criteria for completeness given by the predicates and the subjective assessment of quality expressed by means of a capacity function.

D. Comparison With Axiomatic Approaches

As mentioned in the introduction of this paper, some authors proposed an axiomatic definition of quality measurement. Particularly the proposed definition by Heinrich *et al.* [11] is considered relevant in this paper (see Section II). Therefore, the necessary requirements according to Heinrich *et al.* are reviewed within the scope of our framework.

The first requirement according to [11] is that quality is expressed in a normalized manner, with clear upper and lower bounds. It has been shown in the previous that within our framework, $\mathbb{0}$ and $\mathbb{1}$ are respectively the levels for unacceptable data and data of the highest possible quality. Because the scale $\mathbb{S}$ on which we measure is not fixed, $\mathbb{0}$ and $\mathbb{1}$ set the bounds for a *single* measurement and within *the scope of a particular* database. As a consequence, our framework does not exhibit an *absolute* upper bound of quality. Instead, the upper bound depends on both the measurement and the data. This is considered an important advantage with respect to other approaches. If a database is for example extended by adding attributes, scales can be adjusted to this modification.

The second requirement is that quality is expressed on an interval scale. In our framework, we have loosened this constraint by proposing ordinal scales as it is far from trivial to introduce additivity for several reasons. Except for the case where quality is interpreted as probability (see Section IV-E), Heinrich *et al.* [11] do not detail on the connection between additivity and an empirical concatenation operator [7].

The third requirement stated in [11] is that measurement of quality must be interpretable. The rigorous inference of Q together with several shown properties support the claim that measurement in the proposed framework indeed yields a high degree of interpretation.

As a fourth requirement, it is stated that measurement of quality must be adaptive to a specific setting. This requirement is originally due to Even *et al.* [10], who point out the difference between objective and subjective measurements. In our framework, this is reflected formally as the difference between predicates and capacity. More specifically, the basic assumption on which our theory is built, is that the highest quality of data can be described by a list of criteria, which were then formalized as predicates. It could be argued that an objective measurement would assess any data by requiring that *all* predicates are satisfied. In any other case, the data are not of the highest possible quality and thus not acceptable. However, it is commonly accepted that in a practical usage scenario, data do not have to be of the best quality to be useful. The perceived quality with respect to a specific application is therefore different than the objective requirement that data must be as good as maximally possible. From this perspective, any subjective or application-specific knowledge about data quality is expressed by means of the quality capacity. In other words, the application determines the quality capacity for a given combination of predicates.

The fifth requirement adheres to feasibility and comes down to the fact that it should be practically possible to compute any quality measurement, both in terms of computational complexity and necessary input data. As far as our framework concerns, the definitions of a predicate and a quality capacity are the boundaries in which quality measurements can be defined. However, it is common sense that a predicate that can-not be evaluated with ease, leads to useless measurements.

The sixth and last requirement requires that multiple measurements of data quality can be aggregated to higher-level structures. Heinrich *et al.* [14] and Even *et al.* [8] prescribe the possibility to aggregate measurements of quality on the

attribute level into a measurement of quality on the tuple level. The tuple level can then be aggregated to the relational level and so on. In addition, the interpretation of measurements across all those levels must be the same. This latter requirement is called *interpretation consistency* by Even *et al.* [10], who also prescribe the possibility to aggregate across dimensions. However, this treatment of aggregation makes some assumptions that are hard to advocate. More specifically, aggregation defined in [14] and [8] actually treats two different problems.

The first problem is observed when aggregating from the attribute level to the tuple level or when aggregating across dimensions. Essentially, these are different measurements and the interpretation of their results might also be different, even if the same scale is used. To clarify this, consider the snippet of data from Table I and assume three $[0, 1]$-valued functions f_1, f_2, and f_3 that respectively measure completeness, accuracy, and consistency of an attribute. In this setting, it is clear that a value of 0.8 might carry a different interpretation for each of the three functions. Completeness of 0.8 might be acceptable for the task at hand whereas accuracy of 0.8 renders the data useless. From this perspective, blind aggregation that ignores the meaning of f_1, f_2, and f_3 is virtually without any meaning. The basic problem that we observe here, is that different measurements intrinsically measure something different and therefore the result of different measurements requires separate treatment. Even if these different measurements are expressed in the *same* scale, this does not automatically imply that they can be interpreted in the same way. A combination of different measurements should be interpretation-aware in the sense that the interpretation of each of the individual measurements is taken into account by the aggregation function. Such an interpretation-aware measurement can be obtained by our framework through *recursive* measurements. In such a scheme, the different measurements that must be combined serve as the basis for a more complex quality measurement, which has predicates that rely on the initial measurements and has a capacity that is aware of the interpretation of the individual measurements. As such, a hierarchical system of measures is obtained. We will illustrate this principle in Section VI.

The second problem encountered is the aggregation of quality measurements of different *instances* of data into a global outcome. Suppose that we have a procedure to measure the accuracy of a tuple, then this procedure can be used to measure to quality of all tuples in a relation. All those measurements have a common interpretation as they are obtained with the same measurement procedure. Aggregation of all the tuple measurements is then well defined and provides insight in the quality of the relation. Within our framework, aggregations of this second kind are supported. For each set of data instances $\{R_1, \ldots, R_k\}$, the quality measurements under Q can be aggregated by any aggregation function $a : \mathbb{S}^k \to \mathbb{S}$ that is permitted on an ordinal scale. This means that we can compute minima, maxima, medians, quantiles, etc., to gain insight in the overall quality of the set $\{R_1, \ldots, R_k\}$. For a further reading on aggregation operators, the reader is referred to [39].

An interesting property is that because P is finite so is the image of $\mathcal{C}$. More specifically, the number of possible outcomes of $\mathcal{C}$ is finite and is at most $2^{|P|}$. This means that the quality measurements over the set $\{R_1, \ldots, R_k\}$ can be stored efficiently as a histogram with a worst case space complexity of $2^{|P|}$. In practical scenarios, the space complexity of this histogram will be much lower as is illustrated in Example 1. The advantage of using a histogram representation is that many different aggregations can be computed efficiently because the time complexity of those aggregations reduces to the size of the histogram. In addition, the histogram representation can be updated incrementally, which means that aggregations can be easily distributed over different threads, processes, and machines. In addition, data modifications do not require that the histogram is recalculated.

E. Beyond Ordinal Measurement

In the framework presented so far, measurement is approached in an ordinal manner. It has been shown that such an approach yields an unambiguous framework in which each measurement of quality has a clear interpretation. In this section, the framework is further refined by investigating how expressions that are beyond ordinal scales can be incorporated and how they must be interpreted.

In order to understand the role of scales that are beyond the ordinal level, let us reconsider the concept of "measurement of data quality." As argued in [11], when measuring quality of data, one has in the very essence two options: either one performs a real-world test or one *estimates* the quality. Although this is a very simple and straightforward observation, it is at the same time crucial in our reasoning and argumentation. In case of a real-world test, one simple verifies whether data are a correct representation of reality or not. Note that such a real-world test is a Boolean matter (either data correspond to reality or not) and a real-world test is therefore nothing more than a predicate. In the following, some strategies for performing a real-world test will be discussed and it is in general the best measurement of quality one can perform. However, real-world testing may be not desired (e.g., due to the high cost of it) or simply not possible (e.g., if there is no access to the real-world value). For this reason, the second scenario where quality is estimated will apply in most cases. The question is then *how* quality can be estimated, taking into account that there are many possible causes for degradation of quality (see for example the taxonomy presented in [4]). Despite this complex web of data quality issues, it is motivated here that there are two sensible ways of making an estimation of quality. The first way is to adopt a set of rules or constraints to which data must adhere. Measurement then follows from verification of these rules. This scenario is immediately applicable in the measurement of completeness and consistency and it is a straightforward application of the framework proposed so far. The second way is to adopt a model that describes the uncertainty about the reality and derive an estimation of quality based on this model. Following this method, measuring the quality of data involves measuring the (un)certainty that data are indeed of the highest quality. The nature of the measurement then becomes dependent on the uncertainty theory that is used. Usage of a probabilistic model implies that the corresponding

measurement is quantitative whereas a possibilistic model can result in both quantitative and qualitative measurements.

To support the statement that a quantitative measurement of quality necessarily stems from a model of uncertainty, it is pointed out that literature on data quality actually provides many arguments that back this hypothesis. In [40], Goodchild *et al.* point out that in the context of spatial databases, a measurement of accuracy is closely related to the error model of the instrument that generated the data, which are in many circumstances Gaussian. In [14], it is argued that the currency of data should be measured as the *probability* that data are still up-to-date. More specifically, it is argued there that if shelf life of data assumed to be exponentially distributed, currency of data is given by $\exp(-\eta \cdot (\text{age}(a)))$ where η represents a decline factor [14]. What is of particular interest here is that the probabilistic interpretation of quality is found to be the most sensible way of interpreting quality [11].

If we closer revise the interpretation of measured quality as a quantification of uncertainty, it can be seen that this approach in fact puts forward a Boolean assumption on the connection between data and reality (i.e., a predicate) and then tries to estimate the certainty that this assumption is true. In the probabilistic scenario, this will provide us with a probability that the predicate is true. By this observation, it becomes apparent that there is a close connection between the ordinal framework presented in the previous and the role of "beyond-ordinal" measurement as a way to express uncertainty. Indeed, the presented framework is a measurement framework in which it is assumed that each predicate can be evaluated in a precise manner. Information beyond the ordinal level naturally comes into play when this assumption no longer holds and uncertainty about whether a predicate is true or false needs to be modeled. Within the context of this paper, we present two ways of dealing with this uncertainty: *predicate decomposition* and *uncertainty propagation*.

The first solution is to decompose a predicate with an uncertain evaluation into a chain of predicates that each perform a test on the certainty that the main predicate is true. Suppose a predicate p for which we know the probability distribution $\Pr[p(R) = T]$. Assume that the scale of measurement is given by $\mathbb{S} = \{s_0, \ldots, s_k\}$ with $s_0 = \mathbb{0}$, $s_k = \mathbb{1}$, and $s_i < s_{i+1}$. Under this assumption, we can consider a $(k+1)$-dimensional vector $\mathbf{v} = [v_0, \ldots, v_k] \in [0,1]^k$ such that $v_0 = 0$, $v_k = 1$, and $v_i \leq v_{i+1}$. The $\mathbf{v}$-decomposition of the predicate p is then given by a set of predicates $P^{(\mathbf{v})} = \{p_1, \ldots, p_k\}$ such that

$$\forall i \in \{0, \ldots, k\} : p_i \stackrel{\triangle}{=} \Pr[p(R) = T] \geq v_i. \tag{30}$$

It can be seen that $P^{(\mathbf{v})}$ is a set of k predicates that can be evaluated without uncertainty. Because of the connection between $\mathbf{v}$ and $\mathbb{S}$ by definition, the construction of capacity follows from the consideration $p_i \prec p_{i+1}$.

By predicate decomposition, one immediately translates uncertainty into an appreciation on the scale $\mathbb{S}$ and a decision about reflected quality is taken at the level of predicates. In certain scenarios of decision making, this may not be a desired property. Therefore, a second solution is to propagate the uncertainty about predicates throughout the measurement process in such a way that the outcome of a measurement is in fact an uncertain matter. The result of a measurement is then no longer a value $s \in \mathbb{S}$, but rather a distribution or density function over $\mathbb{S}$. Within the scope of this paper, we limit ourselves to a quick sketch of how such a distribution can be derived. Basically, Theorem 1 serves as the basis for propagation. This theorem implies that we can represent the necessary and sufficient constraints to obtain a given level $s \in \mathbb{S}$ by its minimal and sufficient generator. Taking into account that 1) the capacity function $\mathcal{C}$ is monotonic and 2) $\mathbb{S}$ is a totally ordered set, these minimal and sufficient generators imply a complementary cumulative distribution. Let us clarify this in the case of stochastic uncertainty about predicates. Consider the predicates P and the capacity $\mathcal{C}$ and consider for each $p_i \in P$ the density function X_i as follows:

$$X_i(R) = \Pr[p_i(R) = T]. \tag{31}$$

Under these circumstances, for any $s \in \mathbb{S}$, we can calculate the survival function as

$$\Pr[Q(R) \geq s] = \Pr[G(R \mid s) = T] \tag{32}$$

from which a distribution on $\mathbb{S}$ can be derived. Note that in the specific case where all X_i are distributed mutually independent, we have

$$\Pr[G(R \mid s) = T] = \bigoplus_{B \in \mathcal{B}(s)} \left(\prod_{p_i \in B} X_i(R) \right) \tag{33}$$

where the operator $\oplus$ denotes the probabilistic t-conorm [41]. With these formulae, propagation of uncertainty can be immediately derived. With the above presented results in place, a clear and unambiguous interpretation of assessment of quality beyond the ordinal level has been established. In the following section, the complete framework will be applied on the most important dimensions of data quality.

V. Toward Concrete Measurements

Thus far, an elaborate theoretical exposition of a novel framework for data quality measurement is given. In this section, the framework is applied to obtain measurements for a series of well known quality dimensions. Hereby, dimensions that focus on quality of data as well as on metadata are considered. After a brief and quick discussion of a wide range of dimensions, a more detailed discussion is given about the dimension of consistency.

A. Quick Overview

Perhaps the most important dimension of data quality is the *correctness* of the data. In simple terms, data are correct if it is a true representation of reality. As was mentioned in Section IV-E, measurement of correctness ideally relies on a real-world test. Although this seems infeasible in many cases, there are some ways to establish real-world testing. A first technique is the usage of reference data. Hereby, there is a set of reference data that is curated and maintained in such a way one can assume that the reference data are completely correct at all times. The measurement of data (e.g., a tuple or an attribute value) is performed by cross-checking the data against the

reference data. Hereby, there is a single predicate that asserts that the data are coherent with the reference data. The capacity for such a single predicate is trivial. This technique is often used in measuring the correctness of address data. A second technique is to adopt crowd curation, where a group of experts (i.e., the crowd) is employed to find out if data are correct or not. Hereby, again a single predicate is put forward that asserts correctness of the data. The difference in opinions between experts may lead to uncertainty about the truth of the predicate. As shown in the previous, the proposed framework offers the machinery to deal with such uncertainty.

Two important dimensions of data quality that have been already discussed in the previous are *currency* and *completeness*. With respect to currency, it has been pointed out that the extensive and ample work of Heinrich *et al.* [11], [14], [22] has led to the interpretation of currency as the probability that data are still up-to-date. As shown in the previous, such probability can be integrated in the presented framework. With respect to the dimension of *completeness* of the data, it has been illustrated in Example 1 and further explained in Section IV-D that the proposed framework offers a better handling of this dimension than existing approaches.

Besides the principal dimensions of data quality (including consistency, which is treated separately in the following), there are a number of other dimensions that are frequently used, be it to a lesser extent than the ones already mentioned. Examples are interpretability, believability, and accessibility of data. For each of these dimensions, one can either provide a concrete set of criteria (i.e., predicates) or a model of uncertainty that describes the perceived quality of a piece of data. For interpretability, predicates may include the availability of a database schema, presence of metadata and information on the lineage and/or provenance of the data [6]. For accessibility of the data, predicates may include the presence of multiple representations (e.g., text-based, visual, and auditive), representation of content in multiple languages but also the fact whether or not data can be represented in a device-independent manner [6].

Although the proposed framework is initially intended to measure quality of data, it can also be used to measure quality of metadata. Quality aspects of database design such as schema completeness, schema readability, and schema minimization can be measured with the proposed framework. In addition, there are several well known procedures for measuring or expressing quality on a metalevel that fit into the proposed framework. Perhaps the most obvious example is database normalization. With this procedure, the different normal forms can be seen as levels on the scale $\mathbb{S}$. The measurement of the normal form of a database is done by verifying a set of rules (i.e., predicates) in a certain order (i.e., the measurement is $\prec$-sensitive). Another example is the five-star data paradigm, which uses a five-level scale $\mathbb{S}$ to express the extent to which data are coherent with the linked open data principle. Each level corresponds to certain conditions that must be met, which can be implemented as predicates and a suitable capacity. Finally, a last example of metadata quality is the usage of an error-model to communicate the global accuracy of values. For example, if an attribute "pressure" is measured with a pressure gauge, the error-model of that instrument provides us with an indication of accuracy of all values for that attribute.

B. Measurement of Consistency

An important dimension of data quality that is now discussed in more detail, is *consistency*. According to the definition in [6], consistency "*captures the violation of semantic rules defined over data items*." It can be easily seen that such "semantic rules" are immediately transferable to a set of predicates P. The degradation of quality upon failure of rules can be modeled by means of a suitable capacity function. In order to further detail this, we distinguish three types of consistency: attribute consistency, tuple consistency, and source consistency.

In the case of *attribute consistency*, the scope of the data to which measurement is applied, is that of single attributes. With this kind of measurement, attributes are considered as mutually independent pieces of data and measurement is focused on the adherence of data to certain attribute-specific rules. For this reason, this kind of consistency is sometimes referred to as *conformity*. For an attribute a, the attribute-specific rules induce a set of k nested sets $N_1 \subset N_2 \subset \cdots \subset N_k \subseteq \mathrm{dom}(a)$ in such a way that we can define k predicates as follows:

$$\forall i \in \{1, \ldots, k\} : p_i\left(t[a]\right) \stackrel{\triangle}{=} t[a] \in N_{k-i+1}. \tag{34}$$

This set of predicates can then be measured on the basis of a $\prec$-specific capacity such that $p_i \prec p_{i+1}$. As is pointed out in [13], a compact representation of these nested sets can be obtained in the case of textual attributes. More specifically, when dealing with textual attributes, one may consider a set of predicates $P = \{p_1, \ldots, p_n\}$ in such a way that each predicate p_i either verifies a regular expression pattern Σ (type-I) or verifies a constraint on one or more captured groups of a pattern Σ that was the subject of a predicate p_j with $j < i$. It can be shown that relying solely on these two types of predicates, a simple yet powerful system to define attribute consistency measures is obtained. Currently, about 600 different measurement functions for attribute consistency are defined according to the rules of the framework introduced in this paper. The names of attributes and measures are standardized by means of uniform resource indicators in order to promote and facilitate the exchange of knowledge w.r.t. consistency measurement.

In the case of *tuple consistency*, measurement is done on entire tuples of a relation. Rather than considering single attributes independently of each other, it is recognized that there may exist *dependencies* between attributes. These dependencies can be used to verify whether the co-occurrence of multiple attribute values within the same tuple is in adherence to the rules of consistency. Both functional and inclusion dependencies can play an important role here. Measures of this type can be based on reference data. For example, a registry of first names of newborns per year of birth and gender can be used to measure consistency between attributes "first name," "gender," and "birth date." However, such reference data might be unavailable or hard to maintain. Therefore, measurement of tuple consistency is especially powerful when it can be denoted in a functional form. For example, for Belgian SSNs (an 11-digit identifier

of inhabitants), the first six digits refer to the birth date and the ninth digit is even for females and odd for males. What is important here is that this information does not only inform us about the functional dependencies SSN → Gender and SSN → Birth date, but also informs us about *how* the values for attributes "gender" and "birth date" can be inferred from a given SSN.

The last type of consistency distinguished here is called *source consistency* or source compatibility. In this case, measurement aims to grasp the mutual compatibility of information about a real-world object that originates from two or more different sources. In essence, source consistency can be seen as a special kind of tuple consistency where we encounter the same information more than once. Basically, source consistency handles the same kind of problems encountered also in the field of duplicate detection. From this field, there are many models, both probabilistic and possibilistic, that can be used to quantify the uncertainty about the assertion that two pieces of data are describing the same real-world entity or fact [42], [43]. Source consistency is of particular interest when data are either textual or have a multivalued nature. Relevant examples of multivalued attributes are sets (e.g., hobbies of a person), interval data (e.g., time intervals), vague data (e.g., linguistic labels for age categories), or uncertain data (e.g., ill-known values).

VI. Use-Case: Clinical Trials

In this section, the design and application of quality measures is illustrated in the context of clinical trial data. It is stressed that the purpose of this section is to illustrate the framework in a real-world setting. For indications on the computational complexity of the framework, the reader is referred to [13]. We will report measurement results in this section, but is emphasized that they mainly serve the purpose of showing how the outcome of quality measurement can be interpreted.

A clinical trial is basically a clinical study performed with human participants and is considered as an important aspect in the development of new pharmaceuticals. The aim of clinical trials is to answer one or more specific research questions, for example, the efficacy of a drug in the treatment of a medical condition. Because of their great importance, there are several initiatives that aim to register and provide information about clinical trials. These initiatives may vary in geographical scope (e.g., nation-wide, continental, world-wide) and study purpose (e.g., a collection of trials may be limited to certain medical conditions). Due to this variability, information about a single study is often dispersed across different (independent) databases. As is usual is such a case, maintaining consistency between these databases is a tedious task. In this section, the proposed framework is applied to measure consistency of data coming from two databases. The first database is the European Union Clinical Trial Register (EUCTR) facilitated by the European Medicines Agency. It contains mainly clinical trials from within the European Union.[2] The second database is the Clinical Trials database facilitated by the National Library of Medicine in the U.S. (CTGOV). It contains clinical studies from around the world.[3] For both databases, either the main sponsor or the principal investigator of the study is responsible for inputting and maintaining the necessary data of their studies.

From these two databases, a dataset was created by searching for studies conducted within Belgium and first reported in the database after the first of January, 2010. This query results in 2327 studies in the EUCTR database and 4031 studies in the CTGOV database. The studies from both databases are then linked by their EudraCT number (an identifier available in both databases) which leads to a dataset of 1063 studies. For each of these studies, there are approximately 200 attributes and it is clear that a complete assessment of all these attributes lies outside the scope of the current paper. Instead, we will focus on a specific aspect of the data in order to illustrate the versatility of the consistency dimension and the ability of the proposed framework to account for this versatility and to enable an unambiguous interpretation.

The aspect of the data that is investigated covers the design parameters of a study. These parameters provide a design model of the study and indicate for example whether control groups are used, how test subjects are assigned to groups, etc. In the EUCTR database, information about study design is modeled as a set of Boolean (i.e., "Yes/No") parameters while the CTGOV database uses categorical attributes to model the information.

The analysis of the study design parameters starts with measurement of attribute consistency. The measures that are used for this adopt a three-valued scale $\mathbb{S} = \{0, 1, 2\}$ and rely on two predicates. Predicate p_1 asserts that an attribute value is not NULL and predicate p_2 asserts that an attribute value matches a regular expression pattern. The capacity of this measure is $\prec$-sensitive with $p_1 \prec p_2$. The patterns to verify the design parameters simply list allowed values. For example, pattern "Yes |No" checks if a character string is one of "Yes" or "No" and can be used to verify the Boolean parameters in the EUCTR database. In total, nine measures of this type are applied on the data. The results show that the main degradation of attribute consistency is the occurrence of NULL-values. In addition, it was found that the CTGOV database contains 33 tuples with a deviating spelling for the attribute that indicates the masking of the study (i.e., double blind, single blind, or open label). In general, it can be concluded that study design parameters have high attribute consistency in both databases. More interestingly to discuss is the compatibility between information about study design in both databases. Therefore, the following measure design is considered to measure tuple consistency. The scope in which we measure is a tuple with two attributes: the determinant and the inferred attribute. We adopt a five-level scale $\mathbb{S} = \{0, 1, 2, 3, 4\}$ and define four predicates. The first predicate verifies whether the tuple on which measurement takes place, is not NULL. The second predicate measures attribute consistency of the determinant attribute and verifies if the measured quality is sufficient. This is an example of recursive measurement as was discussed in Section IV-D. As explained in the previous paragraph, attribute consistency of the study design parameters is measured on a three-level scale (i.e., $\{0, 1, 2\}$). In our current measure design, we require the determinant attribute to have attribute consistency of level 2. The third predicate similarly evaluates attribute

[2]https://www.clinicaltrialsregister.eu
[3]http://clinicaltrials.gov

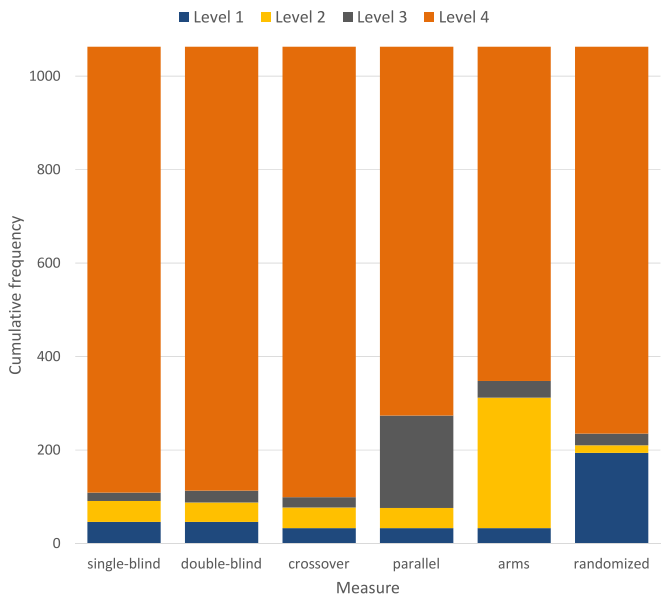

Fig. 1. Tuple consistency results for six measurements based on the dependencies CTGOV:masking → EUCTR:single-blind, CTGOV:masking → EUCTR:double-blind, CTGOV:intervention → EUCTR:crossover, CTGOV:intervention → EUCTR:parallel-groups, CTGOV:intervention → EUCTR:arms, CTGOV:allocation → EUCTR:randomized.

consistency of the inferred attribute. Finally, the fourth predicate evaluates whether the determinant and inferred attribute are in adherence to one another. For example, if the attribute "masking" in the CTGOV database indicates "Double blind," then the Boolean attribute "double blind" in the EUCTR database must have the value "Yes." If the attribute "masking" has another value, attribute "double blind" must have the value "No." As another example, if the attribute "intervention model" (CTGOV) has the value "Single Group Assignment," then the number of arms in the study (EUCTR) must equal 1. For other intervention models, this value must be larger than 1. The capacity of the measure is $\prec$-sensitive in such a way that $p_1 \prec p_2 \prec p_3 \prec p_4$. Following this design pattern, six measures can be defined that all verify a certain aspect of tuple consistency.

The results of applying the six measures to the dataset are shown in Fig. 1. Interpretation of these results is as follows. Tuples with quality 1 are tuples for which the determinant attribute is not of sufficient attribute consistency. In this case, this is always an attribute from the CTGOV database. Tuples with quality 2 have sufficient attribute consistency for the determinant attribute, but not for their inferred attribute (i.e., an EUCTR attribute). By analysis of the attribute consistency, we know that this insufficiency is mainly because of missing data, apart for some spelling variants in the attribute "masking." Tuples with quality 3 are tuples for which attribute consistency of both attributes is sufficient, but there is an inconsistency between the values of the determinant attribute and the inferred attribute. It can be observed that especially for the attribute "parallel-groups" in the EUCTR database, there are quite some conflicts between attribute values. Finally, tuples with quality 4 are considered to have perfect quality w.r.t. the measured aspect.

In their turn, the six measures for tuple consistency can again be used as the basis for new predicates, thus obtaining a global measure of consistency that provides an aggregated view for a single tuple. By doing so, it can be found that 432 tuples (40.6%) that have perfect consistency and 428 tuples (40.3%) have degraded attribute consistency. The remainder of the tuples (19.1%) has perfect attribute consistency but show issues with tuple consistency. For 41 tuples (3.9%), there is more than one issue (i.e., there is more than one tuple consistency measure that results in insufficient quality). It can be seen that the global consistency measure reflects an aggregation strategy that accounts for interpretation of the results that are aggregated and, therefore, maintains disambiguation on an aggregated level.

VII. Conclusion

In this paper, a theoretic framework for the measurement of data quality has been proposed. In contrast to current approaches, the construction of this framework relies on a formal treatment of the concept "measurement." Basically, measurement of quality relies on two corner stones: A set of predicates that can be evaluated against data and a capacity function that expresses the contribution of each combination of predicates with respect to the overall quality. Measurement of quality is then obtained by resolving an ordinal integral. Because of the Boolean nature of predicates, this integral can be reduced to a simple and elegant formula. It is shown that the framework has appealing properties with respect to interpretation and representation. A comparison with other approaches for describing data quality is made and it is shown that our framework has some benefits over these approaches. The aspect of interpretation is further examined in two ways. First, the concept of aggregation of measurements is studied, showing the difference between aggregation of the same measurements on different data and aggregation of different measurements. Second, it is shown how uncertainty about predicates can be integrated in the described process of measurement, thereby linking our framework to existing interval-scaled measures. Finally, the applicability of the framework is illustrated by revising the most important dimensions of data quality and by demonstrating the framework in a real-life use-case.

References

[1] R. Wang, V. Storey, and C. Firth, "A framework for analysis of data quality research," *IEEE Trans. Knowl. Data Eng.*, vol. 7, no. 4, pp. 623–640, Aug. 1995.
[2] C. Batini, C. Cappiello, C. Francalanci, and A. Maurino, "Methodologies for data quality assessment and improvement," *ACM Comput. Surveys*, vol. 41, no. 3, pp. 16–52, 2009.
[3] L. Pipino, Y. Lee, and R. Wang, "Data quality assessment," *Commun. ACM*, vol. 45, no. 4, pp. 211–218, 2002.
[4] W. Kim, E.-K. Hong, S.-K. Kim, and D. Lee, "A taxonomy of dirty data," *Data Mining Knowl. Discovery*, vol. 7, pp. 81–99, 2003.
[5] T. Redman, *Data Quality for the Information Age*. Norwood, MA, USA: Artech House, 1996.
[6] C. Batini and M. Scannapieca, *Data Quality: Concepts, Methodologies and Techniques*. Berlin, Germany: Springer-Verlag, 2006.
[7] D. Krantz, D. Luce, P. Suppes, and A. Tversky, *Foundations of Measurement Volume I: Additive and Polynomial Representations*. New York, NY, USA: Academic, 1971.

[8] A. Even and G. Shankaranarayanan, "Value-driven data quality assessment," in *Proc. Int. Conf. Inf. Quality*, 2005, pp. 265–279.
[9] A. Even and G. Shankaranarayanan, "Understanding impartial versus utility-driven quality assessment in large data-sets," in *Proc. Int. Conf. Inf. Quality*, 2007, pp. 265–279.
[10] A. Even and G. Shankaranarayanan, "Utility-driven assessment of data quality," *Database Adv. Inf. Syst.—ACM SIGMIS J.*, vol. 38, no. 2, pp. 75–93, 2007.
[11] B. Heinrich, M. Klier, and M. Kaiser, "A procedure to develop metrics for currency and its application in CRM," *ACM J. Data Inf. Quality*, vol. 1, no. 1, pp. 5:1–5:28, 2007.
[12] T. Haegemans, M. Snoeck, and W. Lemahieu, "Towards a precise definition of data accuracy and a justification for its measure," in *Proc. Int. Conf. Inf. Quality*, 2016, pp. 16:1–16:13.
[13] A. Bronselaer, J. Nielandt, R. De Mol, and G. De Tré, "Ordinal assessment of data consistency based on regular expressions," in *Proc. Inf. Process. Manage. Uncertainity Knowl.-Based Syst.*, 2016, pp. 317–328.
[14] B. Heinrich, M. Kaiser, and M. Klier, "How to measure data quality? A metric based approach," in *Proc. Int. Conf. Inf. Syst.*, 2007, pp. 1–15.
[15] F. Naumann, J.-C. Freytag, and U. Lesser, "Completeness of integrated information sources," *Inf. Syst.*, vol. 29, no. 7, pp. 583–615, 2004.
[16] C. Fürber and M. Hepp, "Towards a vocabulary for data quality management in semantic web architectures," in *Proc. 1st Int. Workshop Linked Web Data Manage*, 2011, pp. 265–279.
[17] B. Heinrich, M. Kaiser, and M. Klier, "Does the EU insurance mediation directive help to improve data quality?—A metric-based analysis," in *Proc. Eur. Conf. Inf. Syst.*, 2008, pp. 1871–1882.
[18] R. Blake and P. Mangiameli, "The effects and interactions of data quality and problem complexity on classification," *J. Data Inf. Quality*, vol. 2, no. 2, pp. 8:1–8:28, 2011.
[19] S. Stevens, "On the theory of scales of measurement," *Science*, vol. 103, no. 2648, pp. 677–680, 1946.
[20] R. Wang and D. Strong, "Beyond accuracy: What data quality means to data consumers," *J. Manage. Inf. Syst.*, vol. 12, no. 4, pp. 5–34, 1996.
[21] J. Debattista, C. Lange, and S. Auer, "daQ, an ontology for dataset quality information," *Proc. Linked Data Web*, 2014, pp. 1–8.
[22] B. Heinrich and M. Klier, "Metric-based data quality assessment—Developing and evaluation a probability-based currency metric," *Decis. Support Syst.*, vol. 72, pp. 82–96, 2015.
[23] S. d. F. M. Sampaio, C. Dong, and P. Sampaio, "DQ2s—A framework for data quality-aware information management," *Expert Syst. Appl.*, vol. 42, no. 21, pp. 8304–8326, Nov. 30, 2015.
[24] D. Ballou and H. Pazer, "Designing information systems to optimize the accuracy-timeliness tradeoff," *Inf. Syst. Res.*, vol. 6, no. 1, pp. 51–72, 1995.
[25] D. Ballou and H. Pazer, "Modeling completeness versus consistency tradeoffs in information decision systems," *IEEE Trans. Knowl. Data Eng.*, vol. 15, no. 1, pp. 240–243, 2003.
[26] D. Ballou, R. Wang, H. Pazer, and G. K. Tayi, "Modeling information manufacturing systems to determine information product quality," *Manage. Sci.*, vol. 44, no. 4, pp. 462–484, 1998.
[27] D. P. Ballou and G. K. Tayi, "Enhancing data quality in data warehouse environments," *Commun. ACM*, vol. 42, no. 1, pp. 73–78, 1999.
[28] I. Caballero, M. Serrano, and M. Piattini, "A data quality in use model for big data," in *Advances in Conceptual Modeling* (Lecture Notes in Computer Science), M. Indulska and S. Purao, Eds. Berlin, Germany: Springer-Verlag, vol. 8823, 2014, pp. 65–74.
[29] Z. Abedjan, L. Golab, and F. Naumann, "Profiling relational data: A survey," *VLDB J.*, vol. 24, no. 4, pp. 557–581, 2015.
[30] Q. Chen, Z. Tan, C. He, C. Sha, and W. Wang, "Repairing functional dependency violations in distributed data," in *Database Systems for Advanced Applications, PT1* (Lecture Notes in Computer Science), M. Renz, C. Shahabi, X. Zhou, and M. Cheema, Eds. Berlin, Germany: Springer-Verlag, vol. 9049, 2015, pp. 441–457.
[31] G. Cong, F. Wenfei, F. Geerts, X. Jia, and S. Ma, "Improving data quality: Consistency and accuracy," in *Proc. Int. Conf. Very Large Data Base Conf.*, 2007, pp. 315–326.
[32] E. F. Codd, "A relational model of data for large shared data banks," *Commun. ACM*, vol. 13, no. 6, pp. 377–387, 1970.
[33] G. Choquet, "Theory of capacities," *Annales de l'Institut Fourier*, vol. 5, pp. 131–295, 1953.
[34] D. Dubois and H. Prade, *Fundamentals of Fuzzy Sets*. Norwell, MA, USA: Kluwer, 2000.
[35] M. Sugeno, "Theory of fuzzy integrals and its applications," Ph.D. dissertation, Tokyo Inst. Technol., Tokyo, Japan, 1974.
[36] R. Yager, "On ordered weighted averaging aggregation operators in multicriteria decision making," *IEEE Trans. Syst., Man Cybern.*, vol. 18, pp. 183–190, Jan./Feb. 1988.
[37] E. Pap, *Null-Additive Set Functions*. Norwell, MA, USA: Kluwer, 1995.
[38] E. Codd, "Recent investigations in relational data base systems," in *Proc. Int. Fed. Inf. Process. Congr.*, 1974, pp. 1017–1021.
[39] M. Grabisch, J.-L. Marichal, R. Mesiar, and E. Pap, *Aggregation Functions*. Cambridge, U.K.: Cambridge Univ. Press, 2009.
[40] M. Goodchild and K. Clarke, "Data quality in massive data sets," in *Handbook of Massive Data Sets*. Dordrecht, The Netherlands: Kluwer, 2002, pp. 643–659.
[41] E. Klement, R. Mesiar, and E. Pap, *Triangular Norms*. Norwell, MA, USA: Kluwer, 2000.
[42] A. Elmagarmid, P. Ipeirotis, and V. Verykios, "Duplicate record detection: A survey," *IEEE Trans. Knowl. Data Eng.*, vol. 19, no. 1, pp. 1–16, Jan. 2007.
[43] A. Bronselaer and G. De Tré, "Properties of possibilistic string comparison," *IEEE Trans. Fuzzy Syst.*, vol. 18, no. 2, pp. 312–325, Apr. 2010.

Antoon Bronselaer received the M.Sc. degree in computer science and the Ph.D. degree in engineering from Ghent University, Ghent, Belgium, in July 2006 and 2010, respectively.

Since October 2006, he has been a Researcher in the Department of Telecommunications and Information Processing in the research unit "Database, Document, and Content Management," Ghent University. His research interests include data quality and data integration.

Robin De Mol received the M.Sc. degree in computer science engineering in 2008. Since then, he has been working toward the Ph.D. degree at the Database, Document, and Content Management Research Group under the supervision of Prof. G. De Tré.

His research interests include flexible querying and uncertainty modeling.

Guy De Tré received the M.Sc. degree in computer science and the Ph.D. degree in engineering from Ghent University, Ghent, Belgium, in July 1994 and June 2000, respectively.

Since October 2004, he has been a Professor of fuzzy information processing in the Department of Telecommunications and Information Processing, Ghent University, where he heads the research unit "Database, Document, and Content Management." His main research interests include the principles and practice of imperfect information handling in information systems.