Towards a Precise Definition of Data Accuracy and a Justification for its Measure

(Research-in-Progress)

TOM HAEGEMANS, MONIQUE SNOECK and WILFRIED LEMAHIEU, KU Leuven

The goal of this study is to investigate definitions and measurement operations in use for the data accuracy dimension of data quality so that this knowledge can be used to serve as a foundation for a precise definition of data accuracy and a justification for its measure. A multidisciplinary scientific literature review was conducted to collect definitions and measurement operations of data accuracy. These definitions and measurement operations were analysed and coded by a content analysis. It was found that there is a reasonable consensus that data accuracy, for both a single data item and multiple data items, is related to one specific notion, namely the magnitude of an error. This study adds to the knowledge base by identifying a notion of data accuracy that is considerably well-agreed upon, such that it can serve as a foundation for a precise definition of and a justification for a measure of data accuracy.

Categories and Subject Descriptors: [**Methods, Concepts, and Tools for Information Quality**]: IQ Concepts, Metrics, Measures, and Models

General Terms: data quality, data accuracy, definition, metric, measure

1. INTRODUCTION

A key characteristic of a well developed scientific discipline is the precise definition of its constructs and the ability to measure them [Torgerson 1958, p. 2]. Correspondingly, it has been acknowledged that "without a precise definition of what is being measured and without a sound justification for the measures themselves, the assessment of data quality will remain an ad hoc process instead of a scientific one" [Pipino et al. 2005, p. 49].

The most important data quality construct, in both theory and practice, is the data accuracy dimension. In practice, this dimension is indicated as the most important by data consumers [Wang and Strong 1996, p. 13; Nelson et al. 2005, p. 217; Moges et al. 2013, p. 50]. At the same time, in the scientific literature, this data quality construct or dimension is considered to be "key" by many studies in the data quality field [Wand and Wang 1996, p. 87], and is also indicated to be "basic", "straightforward" and "obvious" [Ballou and Pazer 1985, p. 153; Redman 2005, p. 21; Wang et al. 1995, p. 350]. Other data quality constructs, like completeness and correctness, have been precisely defined [Wand and Wang 1996] and received a justification for their measures [Pipino et al. 2005]. However, to the best of our knowledge, this has not been the case for data accuracy.

Before a concept can be precisely defined or can receive a justification for its measure, it is advised to analyse the domain of this construct [Churchill 1979, p. 67] so that its conceptual meaning becomes clear [Giese and Cote 2000, p. 2]. Therefore, the central goal of this research is to analyse how the data accuracy construct is currently defined and measured, so that this analysis can serve as a basis for a precise definition and a justification for the measure of data accuracy. This goal is twofold.

On the one hand, we want to investigate the meaning of the data accuracy concept. The meaning of a concept can be examined by analysing the basic notions used in the definitions and measurement operations of the concept [Giese and Cote 2000, p. 3]. This analysis will also allow to uncover whether the accuracy dimension is a one- or multidimensional concept. Accordingly, the first research question (RQ) of this study can be stated as follows:

RQ 1. *Which and how many notions are used in the state of the art to define data accuracy and are underlying the operations to measure data accuracy?*

On the other hand, this knowledge must pave the way for a precise definition and a justification for the measurement of data accuracy. A measure for a concept can be justified by grounding it in the representational measurement theory [Krantz et al. 1971; Suppes et al. 1989; Luce et al. 1990], which is the most widely accepted measurement theory [Hand 1996, p. 449]. This theory builds on the principle that our intuition should be the starting point for measurement [Fenton and Pfleeger 1996, p. 24; Poels 1999, p. 127] and thus requires empirical decidability [Helzner 2012, p. 603; Luce and Narens 1994, p. 211]. The requirement of empirical decidability has certain prerequisites. One of the prerequisites is that everyone should share the same understanding of the construct to be measured. For instance, when measuring the height of a dog, it is important that everyone knows whether the dog should stand on two or four legs, and whether to measure its height starting from the back or the head of the dog. To enable a shared understanding, it is common practice to construct a *(1)* precise and *(2)* well agreed upon model of the concept [Fenton 1994, p. 204]. Such a model can, for example, take form as a precise definition, whereby a definition of a concept is considered to be precise when it contains only and all the meaning a term has [Belnap 1993, p. 119] and is expressed in a formal way.

As the exact meaning of data accuracy is still not specified unanimously, the precision of the definitions cannot be evaluated yet. However, the degree of consensus about the notion(s) and the number of notion(s) that are used to define and measure data accuracy can be used to evaluate the extent to which the knowledge in the state of the art about data accuracy can be used as a basis for the justification of its measure. In response, the second research question of this study is:

RQ 2. *How well agreed upon are the notions that are used in the state of the art to define data accuracy and are underlying the operations to measure data accuracy?*

2. TERMINOLOGY: DEFINITIONS, METRICS, MEASUREMENT OPERATIONS AND NOTIONS

In this work, we are interested in which notions or ideas are used throughout the literature to describe definitions and measurement operations of data accuracy.

In general, definitions of concepts can be categorised according to their level of abstraction. Theoretical definitions are definitions that explain the meaning of a concept in an abstract way. For example, a theoretical definition of accuracy can be formulated as follows: the "closeness of the agreement between the result of a measurement and a true value of the measurand" [Bureau International des Poids et Mesures 2008, p. 35]. This theoretical definition contains the notion of the magnitude of an error because it defines data accuracy as the "closeness". Operational definitions are definitions that explain the meaning of a concept by stating the operations that are required to measure the concept. An example of an operational data accuracy definition is: "accuracy is defined as the ratio between the number of correct values and the total number of values in a database" [Cappiello et al. 2003, p. 84]. The measurement operations that are delineated in an operational definition can also be written in formal language. In this case, the measurement operation is often referred to by the common term 'metric'. For instance, the metric of Redman [2005, p. 29], as shown in Equation 1, describes the same measurement operation as proposed in the operational definition of Cappiello et al. [2003, p. 84]. Both the exemplified operational definition and metric are based on the notion of the occurrence of an error in data.

$$[\ldots]\ accuracy = \frac{number\ of\ fields\ judged\ \text{“}correct\text{”}}{number\ of\ fields\ tested} \quad (1)$$

In this paper the term 'definition' will be used to refer to a theoretical definition and the term 'measurement operation' will be used to refer to a metric or an operational definition. The term 'notion' will be used to refer to an idea about the meaning of data accuracy that is captured in these definitions or measurement operations (e.g. validity, correctness, ...).

3. METHODOLOGY

To answer the research questions, the notions that were used in definitions of data accuracy and underlie the operations to measure data accuracy in the scientific literature were analysed.

First, the literature in multiple scientific disciplines was reviewed in search for publications that contain definitions or measurement operations of data accuracy.

Next, the definitions and measurement operations of data accuracy were extracted from the relevant papers (see Table II in the appendix). While extracting this information, it became apparent that there were some interesting statements about data accuracy that could not be classified as a definition because they did not explain its meaning. Because these statements contained interesting takes on data accuracy or synonyms of data accuracy, these statements were included in the appendix, but were excluded from our analysis. For example, the statement "also: data quality (as opposed to information quality), error rate, correctness, integrity, precision" [Naumann and Rolker 2000, p. 161] was excluded from the analysis because the authors intended to list synonyms of data accuracy (as indicated by the use of the word "also") while not intending to explain its meaning. Furthermore, sometimes the exact same definition appeared in multiple papers of the same authors. If this was the case, in order to keep the original reference, the definition that was proposed first was kept. On top of that, a single publication sometimes contained both an operational definition and a metric describing exactly the same measurement operation. Since we analyse operational definitions and metrics as one group, in these cases, only the metric was included in the analysis to avoid overrepresentation of an author.

Finally, the definitions and metrics were analysed and coded by a content analysis.

4. RESULTS

Table I summarises the analysis of definitions and measurement operations. The definitions and measurement operations were categorised according to:

— the number of data items the definition or measurement operation applies to. For some definitions or measurement operations, it was clear that it applied to the accuracy of a single data item, while other definitions apply to a data set. For some definitions or measurement operations, this was not clearly specified.
— the scientific branch of the source, either *(1)* social or applied sciences or *(2)* natural sciences. The social sciences consists of scientific disciplines such as information systems research [Recker 2013, p. 12], economics, psychology and sociology. The applied sciences consists of disciplines like operations research, engineering and computer science. The natural sciences consists of scientific disciplines such as physics, chemistry and biology. The analysis revealed differences between natural sciences and the two other groups, while no noticeable differences were remarked between social and applied sciences. Thus, the latter were treated as a single group.
— the type of statement (definition or measurement operation).

Table I. The notions and number of notions that are used in the definitions and measurement operations of accuracy, categorised according to the scientific branch in which they were proposed and the number of data items they concern.

Nr. of items	Scientific branch	Definitions (def.)	Measurement operations (MO)
One	Natural	Total def.: 9 Notion(s) in these 9 def.: — Error magnitude (in 9 def.) Nr. of notions per def.: — One (in 9 def.)	None
	Social and applied	Total def.: 9 Notion(s) in these 9 def.: — Error magnitude (in 9 def.) Nr. of notions per def.: — One (in 9 def.)	Total MO: 2 Notion(s) in these 2 MO: — Error magnitude (in 2 MO) Nr. of notions per MO: — One (in 2 MO)
Multiple	Natural	None	Total MO: 4 Notion(s) in these 4 MO: — Error magnitude (in 4 MO) Nr. of notions per MO: — One (in 4 MO)
	Social and applied	Total def.: 11 Notion(s) in these 11 def.: — Error magnitude (in 4 def.) — Validity (in 1 def.) — Reliability (in 1 def.) — Free of error (in 1 def.) — Damage to utility (in 1 def.) — Correctness (in 5 def.) Nr. of notions per def.: — One (in 10 def.) — Three (in 1 def.)	Total MO: 19 Notion(s) in these 19 MO: — Error magnitude (in 3 MO) — Error occurrence (in 16 MO) Nr. of notions per MO: — One (in 19 MO)
Not specified	Natural	Total def.: 1 Notion(s) in this 1 def.: — Error magnitude (in 1 def.) Nr. of notions per def.: — One (in 1 def.)	None
	Social and applied	Total def.: 5 Notion(s) in these 5 def.: — Error magnitude (in 2 def.) — Level of detail (in 1 def.) — Precision (in 1 def.) — Trueness (in 1 def.) — Free of error (in 2 def.) — Accuracy (in 1 def.) Nr. of notions per def.: — One (in 2 def.) — Two (in 3 def.)	None

4.1. RQ 1: Notions and Number of Notions Used to Define and Measure Data Accuracy

In the natural sciences, according to Table I, the definitions and measurement operations all agree that data accuracy is related to only one notion: the magnitude of an error. This notion is also identified as the sole notion by authorities like the International Organization for Standardization (ISO) [ISO 1994] and the Bureau International des Poids et Mesures [Bureau International des Poids et Mesures 2008]. An example of a measurement operation found in the natural sciences for multiple data items which is based on the magnitude of an error is the mean absolute error.

In the social sciences and applied sciences, as shown in the same table, data accuracy is defined by one, two or three notions depending on the investigated definition.

The definitions contain the following notions: the magnitude of an error, correctness, precision, level of detail, validity, reliability, free of error, damage to utility and even accuracy itself. The notions that were found to underlie the operations to measure data accuracy in the social and applied sciences are the occurrence of an error and the magnitude of an error. The measurement operations only contained a single notion. Remarkably, most of the investigated operations to measure data accuracy contain a notion that was not proposed in *any* of the investigated definitions: the occurrence of an error. An example of a measurement operation found in the social and applied sciences for multiple data items which is based on the occurrence of an error is shown in Equation 1.

4.2. RQ 2: Consensus About the Notions

In the natural sciences, all authors agree that accuracy can be defined and measured by one single notion: the magnitude of an error.

In the social and applied sciences, agreement among the authors concerning the definition and measurement operations of data accuracy is seemingly hard to find.

Fortunately, the notions that are used to define data accuracy in the social and applied sciences are, in fact, also well agreed upon. Most of the notions that are used to define data accuracy in the social and applied sciences other than the magnitude of an error can be considered as a vernacular term (e.g. precision, validity, ...) for data accuracy or a consequence of (in)accurate data (e.g. reliability, damage to utility, ...). As will be explained in Section 5.1, vernacular terms and consequences of data accuracy are not the focus of this study. When the vernacular terms for data accuracy and consequences of (in)accurate data are filtered out, the following single notion remains: the magnitude of an error.

However, the measurement of data accuracy in the social and applied sciences is genuinely inconsistent. Table I shows that the measurement operations of data accuracy for multiple data items in the social and applied sciences are not only based on the magnitude of an error, but also on a different notion: the occurrence of an error. As will be discussed in Section 5.2, we suspect that the introduction of this different notion and its equation with data accuracy is based on sophisms rather than on valid arguments. When this notion is omitted, the remaining operations to measure accuracy are also based on the same notion: the magnitude of an error.

5. DISCUSSION

5.1. Definitions of Data Accuracy in the Social and Applied Sciences

In Section 4.2, we mentioned that the notions that are used to define data accuracy in the social and applied sciences, except for the magnitude of an error, can be considered as a vernacular term for data accuracy or a consequence of (in)accurate data. The use of vernacular terms in definitions of concepts is a frequent procedure in the social sciences. That is, in the social sciences, concepts often encompass certain beliefs and attitudes of individual humans. These beliefs and attitudes can only be assessed by asking questions to individual humans. When constructing a questionnaire to collect these user evaluations it is important to use the vernacular language of the respondents [Payne 1951, p. 12]. For example, the beliefs and attitudes of data consumers about the accuracy of their data is measured by asking them whether they think their data is “correct”, “accurate” or “reliable” [see e.g. Lee and Strong 2003, p. 38]. But, in this work, we are not interested in beliefs or attitudes, but in the actual meaning of the accuracy of data. Thus, the notions that are merely a common term for data accuracy or a consequence of accurate data are not the focus of this study and the following single notion remains: the magnitude of an error.

5.2. Measurement Operations of Data Accuracy in the Social and Applied Sciences

Section 4.2 describes that the measurement operations of data accuracy for multiple data items in the social and applied sciences are not only based on the magnitude of an error, but also on a different notion: the occurrence of an error. The introduction of this different notion has two consequences. First, data accuracy in the social and applied sciences is often measured based on a different notion than the notion underlying the definitions. The notion of the occurrence of an error is not mentioned in *any* of the definitions in our multidisciplinary sample. Second, data accuracy in the social and applied sciences is measured based on a different notion when measuring one data item compared to multiple data items. The notion of the occurrence of an error does not underlie *any* of the measurement operations for data accuracy for a single data item across multiple disciplines.

We believe that the notion of the occurrence of an error to measure data accuracy for multiple data items in the social and applied sciences was unjustly introduced and equated with data accuracy because of at least two elements.

The first element is the urge in the social and applied sciences to aggregate accuracy measurements of multiple data items into one statistic so as to be able to express the accuracy of a dataset rather than a single data item. Table I shows that in the social and applied sciences, the combined majority of definitions and measurement operations of data accuracy apply to multiple data items or an unspecified amount of data items. The same table shows that in the natural sciences, the combined majority of definitions and measurement operations apply to one data item. However, it is important that a statistic that aggregates multiple measurements, is based on the same notion as an individual measurement. If not, the statistic does not correspond to our intuition about the construct and the measure on which this statistic is based on does not adhere to the representational measurement theory [Fenton and Pfleeger 1996, p. 24; Poels 1999, p. 127]. Consequently, "we cannot be sure that the decisions we make based on [this statistic] will have the effects we expect" [Fenton and Pfleeger 1996, p. 106]. If a different notion is used for the aggregated measure statistic, this should be made explicit. For example, consider the measurement of the temperature. Just like the measurement of accuracy for multiple data items (e.g. the accuracy of a database), it is also difficult to express the temperature of multiple points in time (e.g. the temperature of last month). But, this does not imply that the temperature of last month should be expressed by counting the number of days that the temperature was exactly the freezing point much like the accuracy of a database should not be expressed by counting the number of data items where the magnitude of an error is 0. In other words, the need to aggregate multiple measurements so that the accuracy of a whole dataset can be measured does not justify to equate accuracy with the notion of the occurrence of an error instead of the magnitude of an error.

The second element that may have contributed to the popularity of the occurrence of an error to measure data accuracy, is the combination of *(1)* the fact that the meaningfulness of the magnitude of an error in data depends on the scale type of this data and *(2)* the observation that the social and applied sciences more often than in the natural sciences use data of which the scale types do not allow to make meaningful data accuracy statements. The most widely known scale types of data are: nominal, ordinal, interval, ratio and absolute [Stevens 1946; Fenton 1994, p. 201; Fenton and Pfleeger 1996, p. 47]. The nominal scale type allows to categorise entities based on classes, without an ordering or without a notion of magnitude. The ordinal scale type represents an empirical relational system which consists of entities that can be ordered according to a certain dimension. However, the numbers only represent the ranking and do not indicate the differences between two values. The interval scale type also allows to

preserve an order and also preserves the difference between entities. Yet, it does not retain the ratios between these entities because there is no known zero element. The ratio scale type (which is not related to the functional ratio form) can be used when the ratios between the objects are known and the zero element is identified. Therefore, it is possible to preserve the order, size, and ratio between the entities. The absolute scale type is the most advanced scale type and can be used when we are able to count the elements that determine the dimension of an entity. In the natural sciences, one is typically interested in the accuracy of the data of measurements, which is often interval, ratio or absolute while in the social and applied sciences one is typically interested in the accuracy of data of all sorts (e.g. names, postal codes, ...), which can also be nominal or ordinal. However, the meaningfulness of statements about data depends on the scale type [Fenton and Pfleeger 1996, p. 47; Roberts 1985, p. 57]. Consider the following example: suppose two ordinal data items: 'strongly agree' (represented by the number 5) and 'agree' (represented by the number 4), and the following statement about these data items: "the difference between the first and second data item is 1". This statement is not meaningful because the scale type of these data items is ordinal and the exact distance between 'strongly agree' and 'agree' is unknown. Likewise, the magnitude of an error in a single data item cannot be expressed when the scale type of the data is nominal or ordinal because the differences between the elements are unknown. The fact that statements about the magnitude of an error in data are not always meaningful makes it tempting, but not justified, to equate the accuracy construct with the notion of the occurrence of an error. In these cases, the quality of the data should be assessed by other, related, data quality dimensions that do not require a particular scale type.

Based on these considerations it can be seen that the elements above are not valid arguments to measure the accuracy of data in the social sciences with the different notion of the occurrence of an error. When this notion is disregarded, the social and applied sciences also reach agreement about one single notion that should underlie data accuracy measurement: the magnitude of an error.

6. CONCLUSION

Before a concept can be precisely defined and its measure can be justified, it is important to have an understanding about the notion(s) that define(s) the concept. Therefore, before it can be decided whether data accuracy is defined by one or multiple notions and which notions define data accuracy, we first need to know which notions are currently used in the state of the art to define and measure the accuracy of data. If there is consensus about these notions, our results can serve as a starting point for a precise definition and measure of data accuracy. In response, two research questions were stated and answered.

First, we answered RQ 1: "which notions and how many are used in the state of the art to define data accuracy and are underlying the operations to measure data accuracy?".

It was found that the definitions and the operations to measure data accuracy describe different notions. Moreover, the notions in the definitions and operations to measure data accuracy are different depending on the branch of science in which the definition or measurement operation was proposed.

In the natural sciences, all the definitions and measurement operations in our sample contain only one notion: the magnitude of an error.

In the social sciences and applied sciences, the definitions and measurement operations of data accuracy contain multiple notions. The definitions in our sample contained one, two or three notions. The following notions could be identified by analysing these definitions: the magnitude of an error, correctness, precision, level of detail, va-

lidity, reliability, free of error, damage to utility and accuracy. The operations to measure data accuracy in our sample were based on one notion. The following notions could be identified by analysing these measurement operations: the magnitude of an error and the occurrence of an error.

Second, we answered RQ 2: "How well agreed-upon are the notions that are used in the state of the art to define data accuracy and are underlying the operations to measure data accuracy?", it appeared that in the natural sciences, the notions that are used in the definitions and underlie the operations to measure data accuracy are coherent and well-agreed upon. Data accuracy is defined and measured by one notion: the magnitude of an error.

For the social and applied sciences the answer to this research question requires more nuance. Despite being seemingly disorganised, these fields also reach considerable consensus on the meaning of data accuracy. The definitions of data accuracy in the social and applied sciences often refer to a notion that is a common term for data accuracy (e.g. correctness) or is a consequence of (in)accurate data (e.g. reliability). When omitting these notions, there is considerable agreement across the definitions that data accuracy can be defined with a single notion: the magnitude of an error. However, when it comes down to measuring data accuracy, the proposed measurement operations are inconsistent. Measurement operations of data accuracy for multiple data items in the social and applied sciences are not only based on the magnitude of an error, but also on a different notion: the occurrence of an error. We suspect that the introduction of this different notion and its equation with data accuracy is based on sophisms rather than on valid arguments. When this notion is omitted, the remaining operations to measure accuracy are also based on the same notion: the magnitude of an error. Thus, if vernacular terms, consequences of (in)accurate data, and the unjustified notion of the occurrence of an error are filtered away from the results, there is also considerable agreement in the social and applied sciences about the notion that can be used to define and measure data accuracy: the magnitude of an error.

Because it is acceptable to define and measure data accuracy with only one notion, our results can serve as a starting point for a precise definition of data accuracy and a justification for its measure.

7. DIRECTIONS FOR FUTURE RESEARCH

In future work, first, the accuracy of a single data item should be precisely defined and justified by the representational measurement theory. This precise definition and measurement justification should be based on the notion of the magnitude of an error. Next, guidelines to aggregate data accuracy measurements for multiple data items should be formulated so that the accuracy of a dataset can be expressed. On the one hand, the aggregation of these measurements should adhere to the representational measurement theory and should therefore be based on the same notion as a single data accuracy measurement. On the other hand, as with any information, the representation of multiple measurements should be fit for its use in a specific context. Thus, when formulating these guidelines, one also has to take a specific application into account.

ACKNOWLEDGMENTS

Part of this research was funded by KBC Group NV.

REFERENCES

Danielle G T Arts, Nicolette F De Keizer, and Gert-Jan Scheffer. 2002. Defining and Improving Data Quality in Medical Registries: A Literature Review, Case Study, and Generic Framework. *Journal of the American Medical Informatics Association* 9, 6 (2002), 600–11.

Barbara A Bailar. 1985. Quality Issues in Measurement. *International Statistical Review* 53, 2 (1985), 123–139.

Donald P Ballou and Harold L Pazer. 1985. Modeling Data and Process Quality in Multi-Input, Multi-Output Information Systems. *Management Science* 31, 2 (1985), 150–162.

Donald P Ballou and Harold L Pazer. 1987. Cost/Quality Tradeoffs for Control Procedures in Information Systems. *Omega* 15, 6 (1987), 509 – 521.

Donald P Ballou and Harold L Pazer. 1995. Designing Information Systems to Optimize the Accuracy-Timeliness Tradeoff. *Information Systems Research* 6, 1 (1995), 51 – 72.

Carlo Batini and Monica Scannapieco. 2006. *Data Quality: Concepts, Methodologies and Techniques*. Springer. 276 pages.

Nuel D Belnap. 1993. On Rigorous Definitions. *Philosophical Studies* 72, February (1993), 115–146.

Philip R Bevington and Keith D Robinson. 2003. *Data Reduction and Error Analysis for the Physical Sciences* (3 ed.). McGraw-Hill. 320 pages.

Matthew Bovee, Rajendra P Srivastava, and Brenda Mak. 2003. A Conceptual Framework and Belief-Function Approach to Assessing Overall Information Quality. *International Journal of Intelligent Systems* 18, 1 (2003), 51–74.

Bureau International des Poids et Mesures. 2008. *Evaluation of Measurement Data: Guide to the Expression of Uncertainty in Measurement*. Technical Report September. 120 pages.

Cinzia Cappiello, Chiara Francalanci, and Barbara Pernici. 2003. Time-Related Factors of Data Quality in Multichannel Information Systems. *Journal of Management Information Systems* 20, 4 (2003), 71 – 91.

Gilbert A Churchill. 1979. A Paradigm for Developing Better Measures of Marketing Constructs. *American Marketing Association* 16, 1 (1979), 64–73.

Nicole DeHoratius and Ananth Raman. 2008. Inventory Record Inaccuracy: An Empirical Analysis. *Management Science* 54, 4 (2008), 627 – 641.

Churchill Eisenhart. 1962. Realistic Evaluation of the Precision and Accuracy of Instrument Calibration Systems. *J. Res. Nat. Bur. Standards* 67, 2 (1962), 161 – 187.

Churchill Eisenhart. 1968. Expression of the Uncertainties of Final Results. *Science* 160, 3833 (1968), 1201–1204.

Larry P English. 1999. *Improving Data Warehouse and Business Information Quality: Methods for Reducing Costs and Increasing Profits*. Wiley. 518 pages.

Martin J Eppler. 2006. *Managing Information Quality: Increasing the Value of Information in Knowledge-Intensive Products and Processes* (2 ed.). Springer. 398 pages.

Adir Even and G Shankaranarayanan. 2007. Utility-Driven Assessment of Data Quality. *ACM SIGMIS Database* 38, 2 (may 2007), 75.

Wenfei Fan and Floris Geerts. 2012. Foundations of Data Quality Management. *Synthesis Lectures on Data Management* 4, 5 (2012), 217.

Norman E Fenton. 1994. Software Measurement: A Necessary Scientific Basis. *IEEE Transactions on Software Engineering* 20, 3 (mar 1994), 199–206.

Norman E Fenton and Shari Lawrence Pfleeger. 1996. *Software Metrics: A Rigorous and Practical Approach* (2 ed.). Thomson Publishing. 638 pages.

Craig W Fisher, Eitel J M Lauria, InduShobha N Chengalur-Smith, and Richard Y Wang. 2011. *Introduction to Information Quality*. AuthorHouse. 254 pages.

Craig W Fisher, Eitel J M Lauria, and Carolyn C Matheus. 2009. An Accuracy Metric: Percentages, Randomness, and Probabilities. *Journal of Data and Information Quality* 1, 3 (2009), 16:1 – 16:21.

Christopher Fox, Anany Levitin, and Thomas C Redman. 1994. The Notion of Data and Its Quality Dimensions. *Information Processing & Management* 30, I (1994), 9–19.

Joan L Giese and Joseph A Cote. 2000. Defining Consumer Satisfaction. *Academy of Marketing Science Review* 1 (2000).

David J Hand. 1996. Statistics and the Theory of Measurement. *Journal of the Royal Statistical Society* 159, 3 (1996), 445 – 492.

Jeffrey Helzner. 2012. On the Representation of Error. *Synthese* 186, 2 (2012), 601–613.

Y U Huh, F R Keller, Thomas C Redman, and A R Watkins. 1990. Data Quality. *Information and Software Technology* 32, 8 (1990), 559–565.

Rob J Hyndman and Anne B Koehler. 2006. Another Look at Measures of Forecast Accuracy. *International Journal of Forecasting* 22, 4 (oct 2006), 679–688.

ISO. 1994. *ISO 5725-1*. Technical Report. International Organization for Standardization.

Matthias Jarke, Maurizio Lenzerini, Yannis Vassiliou, and Panos Vassiliadis. 2003. *Fundamentals of Data Warehouses* (2 ed.). Springer. 219 pages.

David H Krantz, Duncan R Luce, Patrick Suppes, and Amos Tversky. 1971. *Foundations of Measurement: Additive and Polynomial Representations*. Vol. 1. Academic Press. 567 pages.

Yang W Lee, Leo L Pipino, James D Funk, and Richard Y Wang. 2006. *Journey to Data Quality*. The MIT Press.

Yang W Lee and Diane M Strong. 2003. Knowing-Why About Data Processes and Data Quality. *Journal of Management Information Systems* 20, 3 (2003), 13 – 39.

Duncan R Luce, David H Krantz, Patrick Suppes, and Amos Tversky. 1990. *Foundations of Measurement: Representation, Axiomatization and Invariance*. Vol. 3. Academic Press. 356 pages.

Duncan R Luce and Louis Narens. 1994. Fifteen Problems Concerning the Representational Theory of Measurement. *Patrick Suppes: Scientific Philosopher* 2 (1994), 219–249.

Antonio Menditto, Marina Patriarca, and Bertil Magnusson. 2007. Understanding the Meaning of Accuracy, Trueness and Precision. *Accreditation and Quality Assurance* 12, 1 (2007), 45–47.

Jerzy Michnik and Mei-Chen Lo. 2009. The Assessment of the Information Quality With the Aid of Multiple Criteria Analysis. *European Journal of Operational Research* 195, 3 (2009), 850–856.

Holmes Miller. 1996. The Multiple Dimensions of Information Quality. *Information Systems Management* 13, 2 (1996), 79 – 82.

Helen-Tadesse Moges, Karel Dejaeger, Wilfried Lemahieu, and Bart Baesens. 2013. A Multidimensional Analysis of Data Quality for Credit Risk Management: New Insights and Challenges. *Information & Management* 50, 1 (jan 2013), 43–58.

R B Murphy. 1961. On the Meaning of Precision and Accuracy. In *Materials Research and Standards*. Atlantic City.

Felix Naumann and Claudia Rolker. 2000. Assessment Methods for Information Quality Criteria. In *International Conference on Information Quality*. 148–162.

R Ryan Nelson, Peter A Todd, and Barbara H Wixom. 2005. Antecedents of Information and System Quality: An Empirical Examination Within the Context of Data Warehousing. *Journal of Management Information Systems* 21, 4 (2005), 199 – 235.

Jack E Olson. 2003. *Data Quality: The Accuracy Dimension*. Morgan Kaufmann. 293 pages.

Stanley L Payne. 1951. *The Art of Asking Questions*. Princeton University Press.

Leo L Pipino, Richard Y Wang, David Kopcso, and William Rybolt. 2005. Developing Measurement Scales for Data-Quality Dimensions. In *Information Quality*, Richard Y Wang, Elizabeth M Pierce, Stuart E Madnick, and Craig W Fisher (Eds.). M.E. Sharpe, Armonk, NY, Chapter 3, 37 – 51.

Geert Poels. 1999. *On the Formal Aspects of the Measurement of Object-Oriented Software Specifications*. Ph.D. Dissertation. Katholieke Universiteit Leuven.

Semyon G Rabinovich. 2013. *Evaluating Measurement Accuracy: A Practical Approach* (2 ed.). Springer. 1–271 pages.

Jan Recker. 2013. *Scientific Research in Information Systems*. Springer. 150 pages.

Thomas C Redman. 1996. *Data Quality for the Information Age*. Artech House. 303 pages.

Thomas C Redman. 2005. Measuring Data Accuracy: A Framework and Review. In *Information Quality*, Richard Y Wang, Elizabeth M Pierce, Stuart E Madnick, and Craig W Fisher (Eds.). M.E. Sharpe, Armonk, NY, Chapter 2, 21 – 36.

Fred S Roberts. 1985. *Measurement Theory*. Cambridge University Press.

V Sessions and M Valtorta. 2009. Towards a Method for Data Accuracy Assessment Utilizing a Bayesian Network Learning Algorithm. *Journal of Data and Information Quality* 1, 3 (2009), 14:1 – 14:34.

Stanley Smith Stevens. 1946. On the Theory of Scales of Measurement. *Science* 103, 2684 (1946), 677–680.

Patrick Suppes, David H Krantz, Duncan R Luce, and Amos Tversky. 1989. *Foundations of Measurement: Geometrical, Threshold and Probabilistic Representations*. Vol. 2. Academic Press. 493 pages.

Warren S Torgerson. 1958. *Theory and Methods of Scaling*. Wiley.

Yair Wand and Richard Y Wang. 1996. Anchoring Data Quality Dimensions in Ontological Foundations. *Commun. ACM* 39, 11 (nov 1996), 86–95.

Richard Y Wang, M P Reddy, and Henry B Kon. 1995. Toward Quality Data: An Attribute-Based Approach. *Decision Support Systems* 13, 3 (1995), 349–372.

Richard Y Wang and Diane M Strong. 1996. Beyond Accuracy : What Data Quality Means to Data Consumers. *Journal of Management Information Systems* 12, 4 (1996), 5–33.

W J Youden. 1961. How to Evaluate Accuracy. In *Materials Research and Standards*. Atlantic City.

8. APPENDIX

Code	Statement	Source	Kind
1	"the extent to which registered data are in conformity to the truth"	[Arts et al. 2002, p. 603]	D
2	"The accuracy dimension is the most straightforward and is merely the difference between the correct value and that actually used. Differential"	[Ballou and Pazer 1985, p. 153]	M
3	"the recorded value is in conformity with the actual value"	[Ballou and Pazer 1985, p. 153]	D
4	"Addressing the accuracy dimension is straightforward. If the recorded value is not what it should be, the data unit is labelled as defective"	[Ballou and Pazer 1987, p. 514]	S
5	"the recorded value is in conformity with the actual value"	[Ballou and Pazer 1987, p. 513]	D
6	"the degree to which the reported value is in conformance with the actual or true value"	[Ballou and Pazer 1995, p. 52]	D
7	"Semantic accuracy is the closeness of the value v to the true value v'."	[Batini and Scannapieco 2006, p. 21]	D
8	"Accuracy is defined as the closeness between a value v and a value v', considered as the correct representation of the real-life phenomenon that v aims to represent."	[Batini and Scannapieco 2006, p. 20]	D
9	"Syntactic accuracy is the closeness of a value v to the elements of the corresponding definition domain D."	[Batini and Scannapieco 2006, p. 20]	D
10	"The accuracy of an experiment is a measure of how close the result of the experiment is to the true value"	[Bevington and Robinson 2003, p. 2]	D
11	"Refers to information being true or error free with respect to some known, designated, or measured value"	[Bovee et al. 2003, p. 59]	D
12	"Accuracy is defined as the ratio between the number of correct values and the total number of values in a database"	[Cappiello et al. 2003, p. 84]	M
13	"We [. . .] define inventory record inaccuracy as the absolute difference between the recorded and actual inventory quantity [. . .]."	[DeHoratius and Raman 2008, p. 629]	M
14	"Accuracy to surrogate source. A measure of the degree to which data agrees with an original source of data, such as a form, document, or unaltered electronic data received from outside the control of the organisation that is acknowledged to be an authoritative source."	[English 1999, p. 142]	D
15	"Accuracy (to reality). The degree to which data accurately reflects the real-world object or event being described. "	[English 1999, p. 142]	D
16	"Accuracy is the highest degree of inherent information quality possible."	[English 1999, p. 142]	S
17	"Is the information precise enough and close enough to reality?"	[Eppler 2006, p. 8]	S
18	"Degree of conformity of a measure to a standard or a true value. "	[Eppler 2006, p. 364]	D
19	"Level of precision or detail."	[Eppler 2006, p. 364]	D
20	"observed accuracy: "The correctness of data items, compared to a baseline""	[Even and Shankaranarayanan 2007, p. 83]	D
21	"impartial accuracy: "The extent to which the data items included in the dataset are correct""	[Even and Shankaranarayanan 2007, p. 83]	D
22	"contextual accuracy: "The extent to which incorrect data items damage utility""	[Even and Shankaranarayanan 2007, p. 83]	D
23	"We therefore define the data item accuracy as reflecting the extent to which the content of attribute in record is different from a baseline value which is perceived to be correct."	[Even and Shankaranarayanan 2007, p. 86]	D
24	"Data accuracy refers to the closeness of values in a database to the true values of the entities that the data in the database represent"	[Fan and Geerts 2012, p. 4]	D
25	"Accuracy refers to how closely the data specifically represents the real world."	[Fisher et al. 2011, p. 55]	D
26	"Accuracy generally means that the recorded value conforms to the real-world value, and refers to lack of errors or free of errors"	[Fisher et al. 2011, p. 55]	D
27	"Accuracy of a datum refers to the degree of closeness of its value v to some value v' in the attribute domain considered correct for the entity e and the attribute a."	[Fox et al. 1994, p. 14]	D
28	"Accuracy is a measure of agreement with an identified source."	[Huh et al. 1990, p. 560]	D
29	"[T]he validity of the data, with respect to real-world values"	[Jarke et al. 2003, p. 155]	D
30	"The dimension of accuracy itself, however, can consist of one or more variables, only one of which is whether the data is correct"	[Lee et al. 2006, p. 55]	S
31	"Freedom from mistake or error; conformity to truth or to a standard or model; degree of conformity of a measure to a standard or a true value"	[Michnik and Lo 2009, p. 852]	D
32	"Accurate information reflects the underlying reality"	[Miller 1996, p. 79]	S
33	"Quotient of the number of correct values in the source and the overall number of values in the source. "	[Naumann and Rolker 2000, p. 161]	M
34	"Also: data quality (as opposed to information quality), error rate, correctness, integrity, precision"	[Naumann and Rolker 2000, p. 161]	S
35	"[Data accuracy] refers to whether the data values stored for an object are the correct values. To be correct, a data value must be the right value and must be represented in a consistent and unambiguous form."	[Olson 2003, p. 29]	D
36	"Accuracy of a datum $< e, a, v >$ refers to the nearness of the value v to some value v' in the attribute domain, which is considered as the (or maybe only a) correct one for the entity e and the attribute a."	[Redman 1996, p. 255]	D
37	"Accuracy measures the degree of correctness of a given collection of data"	[Redman 2005, p. 24]	D
38	"[. . .] [I]t is easy enough to quantify the inaccuracy, as the difference between the actual and recorded [data]"	[Redman 2005, p. 24]	M
39	"Accuracy is considered how close a measurement, or data record, is to the real-world situation it represents"	[Sessions and Valtorta 2009, p. 3]	D
40	"There is no exact definition for accuracy. In terms of our model we propose that inaccuracy implies that information system represents a real-world state different from the one that should have been represented."	[Wand and Wang 1996, p. 93]	S
41	"The extent to which data are correct, reliable, and certified free of error"	[Wang and Strong 1996, p. 31]	D
42	"Measurement accuracy reflects the closeness between the measurement result and the true value of the measurand. Measuring instruments are created by humans, and every measurement on the whole is an experimental procedure. Therefore, results of measurements cannot be absolutely accurate."	[Rabinovich 2013, p. 2]	D
43	"The word accuracy conveys an idea of being close to the 'truth', "	[Bailar 1985, p. 126]	D
44	"Closeness of agreement between a quantity value obtained by measurement and the true value of the measurand"	[Menditto et al. 2007, p. 45]	D
45	"The closeness of agreement between a test result and the accepted reference value."	[ISO 1994, p. 3.6]	D
46	"[. . .] the degree of agreement of such measurements with the true value of the magnitude of the quantity concerned"	[Eisenhart 1962, p. 172]	D
47	"[. . .] accuracy is determined by the closeness to the true value characteristics of such measurements"	[Eisenhart 1968, p. 1201]	D
48	"Accuracy should connote the idea of the error of individual measurements when that error is compounded of bias or systematic error and random or nonsystematic error"	[Murphy 1961, p. 266]	D
49	"The term accuracy conveys to the most the idea of a value that is very close to the truth"	[Youden 1961, p. 268]	D
50	"closeness of the agreement between the result of a measurement and a true value of the measurand"	[Bureau International des Poids et Mesures 2008, p. 35]	D
51		[Arts et al. 2002, p. 601]	M

$$Inaccuracy = \frac{InaccurateValues}{TotalValues}$$

52 [Batini and Scannapieco 2006, p. 23] M

$$WeakAccuracyError = \sum_{i=1}^{N} \frac{\beta((q_i > 0) \wedge (s_i = 0))}{N}$$

β is a boolean variable equal to 1 if the condition in parentheses is true, 0 otherwise.
$q_{ij} (i = 1 \ldots N, j = 1 \ldots K)$ a boolean variable defined to correspond to the cell values y_{ij} such that q_{ij} is equal to 0 if y_{ij} is syntactically accurate, while otherwise it is equal to 1.
$q_i = \sum_{j=1}^{K} q_{ij}$
s_i is a boolean variable equal to 1 if the value affects identification, 0 otherwise.

53 [Batini and Scannapieco 2006, p. 23] M

$$StrongAccuracyError = \sum_{i=1}^{N} \frac{\beta((q_i > 0) \wedge (s_i = 1))}{N}$$

β is a boolean variable equal to 1 if the condition in parentheses is true, 0 otherwise.
$q_{ij} (i = 1 \ldots N, j = 1 \ldots K)$ a boolean variable defined to correspond to the cell values y_{ij} such that q_{ij} is equal to 0 if y_{ij} is syntactically accurate, while otherwise it is equal to 1.
$q_i = \sum_{j=1}^{K} q_{ij}$
s_i is a boolean variable equal to 1 if the value affects identification, 0 otherwise.

54 [Batini and Scannapieco 2006, p. 23] M

$$DegreeOfSyntacticAccuracy = \sum_{i=1}^{N} \frac{\beta((q_i = 0) \wedge (s_i = 0))}{N}$$

β is a boolean variable equal to 1 if the condition in parentheses is true, 0 otherwise.
$q_{ij} (i = 1 \ldots N, j = 1 \ldots K)$ a boolean variable defined to correspond to the cell values y_{ij} such that q_{ij} is equal to 0 if y_{ij} is syntactically accurate, while otherwise it is equal to 1.
$q_i = \sum_{j=1}^{K} q_{ij}$
s_i is a boolean variable equal to 1 if the value affects identification, 0 otherwise.

55 [Cappiello et al. 2003, p. 84] M

$$AccuracyOfOperationalDatabases_{ij} = Local\_accuracy_{ij} - outofdate_{ij}$$

56 [Even and Shankaranarayanan 2007, p. 86] M

$$a^{E}_{m,n} = dist(f^{E}_{m,n}, {}^{E*}_{m,n})$$

$$dist : D \times D \rightarrow [0;1] : (f^{E}_{m,n}, {}^{E*}_{m,n}) \mapsto \begin{cases} f^{E}_{m,n} = {}^{E*}_{m,n} & 1 \\ f^{E}_{m,n} \neq {}^{E*}_{m,n} & [0;1[\end{cases}$$

57 [Fisher et al. 2009, p. 5] M

$$accuracy = (\frac{NrOfCorrectValues}{TotalNrOfValues}, RandomnessOfTheOccuranceOfAnError, ProbabilityDistributionOfTheOccuranceOfAnError)$$

58 [Fox et al. 1994, p. 14] M

$$AccuracyOfNumericalValues = InaccuracyOfNumericalValues = v' - v$$

59 The authors propose a series of statistics based on the magnitude of an error: mean square error (MSE), root mean square error (RMSE), mean absolute error, mean absolute percentage error, ... [Hyndman and Koehler 2006, p.] M

60 [Lee et al. 2006, p. 55] M

$$Free\text{-}of\text{-}error\ rating = 1 - (\frac{Number\ of\ data\ units\ in\ error}{Total\ number\ of\ data\ units})$$

61 The authors propose a Bayesian network to create association rules to predict the number of incorrect data items. [Sessions and Valtorta 2009, p. 16] M

62 [Redman 2005, p. 29] M

$$field\ level\ accuracy = \frac{number\ of\ fields\ judged\ \text{“}correct\text{”}}{number\ of\ fields\ tested}$$

63 [Redman 2005, p. 29] M

$$record\ level\ accuracy = \frac{number\ of\ records\ judged\ \text{“}completely\ correct\text{”}}{number\ of\ records\ tested}$$

64 [Redman 1996, p. 256] M

$$p = \frac{number\ of\ Number\ of\ correct\ values}{Number\ of\ total\ values}$$

65 “the measure accuracy is an assessment of the percent of records whose values for a given field are accurate as confirmed with its actual values” [English 1999, p. 147] M

66 “We measure accuracy as the percentage of the data of a relation that capture the actual, real-world values of the entities they represent” [Jarke et al. 2003, p. 164] M

67	The author proposes to measure accuracy with the MSE, but warns that this is only a statistic and does not tell the whole story	[Eisenhart 1962, p. 179]	M
68	Root mean square error	[Murphy 1961, p. 360]	M
69	The author proposes to measure the "bias", which is the difference between an observed and a reference level.	[Murphy 1961, p. 360]	M
70	The author proposes to measure the "limits of error", which is the difference between an observed and a reference level plus and minus three standard deviations.	[Murphy 1961, p. 360]	M

Table II: The Extracted Statements (S), Definitions (D) and Measurement Operations (M) of Data Accuracy