xXXx

Yilin Huang, Alexander Verbraeck XX
Delft University of Technology
Faculty of Technology, Policy & Management
Systems Engineering & Simulation

November 13, 2018

**

Keywords: *

1 Introduction

Data Quality Issues Although data quality varies widely from case to case, data is often of poor quality in practice. In this research, data quality is accessed according to how well the data provides information for model component identification, composition and configuration. Common data quality criteria are listed in Table 1, which include accuracy, completeness, consistency, presentation suitability, among others. Detecting and improving data quality issues need domain understanding and data understanding.

1.0.1 Data Quality Criteria

Accuracy. Although many data quality studies include accuracy as a key criterion, there is no commonly accepted definition of what it means exactly[1], in particular its difference to quality criteria such as correctness (Wand and Wang, 1996; Price and Shanks, 2005). We deem data as being accurate when the data values stored in the database are in conformity with the actual or defined values (Ballou and Pazer, 1985; Fox et al., 1994). Moreover, we distinguish syntactic accuracy from semantic accuracy (Scannapieco et al., 2005; Batini et al., 2009). *Syntactic accuracy* is the conformity of a data value v to the corresponding definition domain D of the data value (Batini et al., 2009). *Semantic accuracy* is the conformity of a data value v to its real-world value v' that is considered correct (Fox et al., 1994; Redman, 1996; Batini et al., 2009). For example, that v="Yilin" when v'="Chris" is considered syntactically accurate but semantically inaccurate if the definition domain is specified as character string. Precision as a data quality criterion does not pertain to semantic accuracy according to this definition, as the conformity can be defined to be an approximation, i.e., $v \approx v'$, to tolerate imprecision.

Completeness. Wand and Wang (1996) and Batini et al. (2009) define completeness as the degree to which a given data collection includes data describing the corresponding set of real-world objects and phenomena. Data completeness can be both semantic and pragmatic. *Semantic completeness* is the degree to which existing values are included in a data collection relevant to the purpose for which the data is stored[2] (Redman, 1996; Bovee et al., 2003; Price and Shanks, 2005). Consider a data collection of employees' birth dates and driving license numbers. A `null` value signifies incompleteness for a birth date but not necessarily for a driving license number as not everyone must have a driving license. On the other hand, the non-occurrence of `null` values does not necessarily entail semantic completeness. When, e.g., a new employee's data does not appear in the database at all, the data is incomplete although there is no `null` value that signifies this incompleteness. We may call missing values of this nature "missing record". A large part of data (in-)completeness is related to the purpose of data use. *Pragmatic completeness* is considered in relation with the purpose of data use rather than the (original) purpose for which the

[1] For example, Gelbstein (2003) defines accuracy as the opposite of an error.

[2] Semantic completeness is often related to the `null` values in a database; a `null` value connotes a missing data value (i) that exists but is not known, (ii) that does not exist, or (iii) that is not known whether it exists or not (Redman, 1996; Price and Shanks, 2005; Batini et al., 2009). Only the first case is seen as being semantic incomplete by our definition.

Category	#	Criterion	Definition	Reference	Example/Explanation
A. Syntactics	1.	Syntactic accuracy	The conformity of a data value v to the corresponding definition domain D of the data value.	Scannapieco et al. (2005) and Batini et al. (2009)	Accurate: v="Yilin", v'="Chris" Inaccurate: v="Yilin", v'=2012 for D=`varchar`
	2.	Syntactic consistency	The uniformity in the syntactic representation of data values that have the same or similar semantics.	Pipino et al. (2002) and Loshin (2011)	Inconsistent: in table-1 employee.id=1234 while in table-2 enrollment.id="1234".
B. Semantics	3.	Semantic accuracy	The conformity of a data value v to its real-world value v' that is considered correct.	Fox et al. (1994), Redman (1996), and Batini et al. (2009)	Accurate: v=10.1, v'=10 Inaccurate: v=11, v'=10 for $\|v-v'\| \leq 0.1$
	4.	Semantic completeness	The degree to which existing values are included in data relevant to the purpose for which the data is stored.	Redman (1996), Bovee et al. (2003), and Price and Shanks (2005)	Incomplete: there are missing values in the data.
	5.	Mapping consistency	The uniformity in the key values of data representing the same external instance.	Price and Shanks (2005)	Inconsistent: keys assigned with different values intend to map the same external instance.
C. Pragmatics	6.	Pragmatic completeness	The degree to which data is of sufficient breadth, depth and scope for the purpose of data use.	Wang and Strong (1996)	Incomplete: there is missing information for a given use of the stored data.
	7.	Timeliness	The extent to which data is within a valid time frame with respect to the purpose of data use.	Wang and Strong (1996) and Price and Shanks (2005)	University course schedules are valid for a given time frame (e.g., a particular semester).
	8.	Presentation suitability	The degree to which the data format, unit, precision and type-sufficiency are appropriate for the purpose of data use.	Price and Shanks (2005) and McGilvray (2008)	Data format, unit, precision and type-sufficiency are sub-criteria.
		- *Precision*	The degree to which each data value expresses sufficient detail that is appropriate for the purpose of data use.	Pipino et al. (2002) and Price and Shanks (2005)	The appropriateness of image resolution is dependent on the use of images.
		- *Type-sufficiency*	The degree to which data includes all of the types of information useful for the purpose of data use.	Price and Shanks (2005)	

Table 1: Definitions and examples of data quality criteria

data is collected. Wang and Strong (1996) define this as the degree to which data is of sufficient breadth, depth, and scope for the purpose of data use, i.e., whether there is missing information for a given use of the stored data.

Consistency. Intuitively, data is deemed consistent when there is no contradiction or disagreement in the stored data (ibid.). Data consistency issues can be found in syntactic and semantic categories. *Syntactic consistency* refers to uniformity in the (syntactic) representation of data values that have same or similar semantics (Pipino et al., 2002; Loshin, 2011). This means that data with the same semantics should best share the same underlying syntactic formats and structures. *Semantic consistency* refers to the conformity of (explicit or implicit) semantic rules over a set of data attributes and values (Batini et al., 2009). Ideally, similar data attributes should share consistent names and meanings (Loshin, 2011) and inter-related attribute values should not have conflicting or unaccountable meanings. Price and Shanks (2005) further differentiate consistency in key values with that in non-key values[3]. Semantic consistency in *key values* – let us call it *mapping consistency* – refers to the uniformity in the key values of data representing the same external instance (ibid.). More specifically, when keys assigned with different values indeed intend to map the same external instance, these keys are considered semantically inconsistent although they may be syntactically consistent. Mapping inconsistency often occurs across databases or data repositories. Some authors call it identifiablility or object identification problem (e.g., Batini and Scannapieco, 2006; Loshin, 2011). Semantic consistency in *non-key values*[4] is frequently mentioned in literature, and evidently it often appears in data. Nonetheless, we *exclude* it as a data quality criterion. Inconsistency often occurs in real-world objects and phenomena. It does not necessarily indicate erroneous data values and hence we argue that it is not a valid criterion for data quality. For the data values that do require semantic consistency[5], the data quality is covered by the semantic accuracy criterion since the values can not be accurate when they are not consistent.

The accuracy criteria mentioned earlier concerns both key values and non-key values. Semantic accuracy in non-key values implies semantic accuracy in mapping (i.e., the associated key-values), not vice versa. More specifically, when a non-key value is semantically accurate, it has a meaningful and unambiguous mapping[6].

Timeliness. It may refer to the time expectation for accessibility of data (e.g., Loshin, 2011), the delay between a change of a real-world state and the resulting modification of the information system state (e.g., Wand and Wang, 1996), or how up-to-date the data is with respect to the task it is used for (e.g., Wang and Strong, 1996; Pipino et al., 2002), etc. Some authors (e.g., Fox et al., 1994; Catarci and Scannapieco, 2002; Bovee et al., 2003; Batini and Scannapieco, 2006; Batini et al., 2009) characterize timeliness with sub-criteria such as *currency* (i.e., how recent is the data, or how promptly the data is updated) and *volatility* (i.e., how long the data remains valid, or how frequently the data varies in time). Timeliness in this research is in the pragmatic category and it refers to the extent to which data is within a valid time frame with respect to the purpose of data use (Price and Shanks, 2005).

Presentation Suitability. This criterion is in the pragmatic quality category and it refers to the degree to which the data format, unit, precision and type-sufficiency are appropriate for the purpose of data use (Price and Shanks, 2005; McGilvray, 2008). Data *precision*[7] is defined as the degree to which each data value expresses sufficient detail that is appropriate for the purpose of data use (Pipino et al., 2002; Price and Shanks, 2005); *type-sufficiency* refers to the degree to which the data includes all the types of information useful for the purpose of data use (Price and Shanks, 2005).

To provide an overview, Table 2 numerates the proposed data quality categories and criteria. Table 1 summarizes the definitions of the eight criteria ordered to the three categories.

	Accuracy	**Completeness**	**Consistency**	**Timeliness**	**Presentation Suitability**
(A) Syntactics	#1.	-	#2.	-	-
(B) Semantics	#3.	#4.	#5. (mapping)	-	-
(C) Pragmatics	-	#6.	-	#7.	#8.

Table 2: Proposed data quality categories and criteria

[3]A key (or mapping) value maps a (non-key) data value (or units) to a represented external (e.g., real-world) instance; a non-key value is a representation (of an attribute) of the external instance itself (Price and Shanks, 2005).

[4]For example, when person A's marital status is "married" and person A's spouse is person B, there would be a semantic inconsistency if person B's marital status is "single".

[5]For example, an under-age child can not be "married", neither does the child have a driving license.

[6]A non-key value is meaningfully mapped when it refers to at least one specific external instance; it is unambiguously mapped when it refers to at most one specific external instance (Price and Shanks, 2005).

[7]When precision is related to measurement systems, it is the degree to which repeated measurements under unchanged conditions show the same results (Taylor, 1999). As for data quality, we support the view that data precision should be considered with respect to data use (Fox et al., 1994; Levitin and Redman, 1995; Price and Shanks, 2005). Data precision without context is often meaningless. After all, no real measurement is infinitely precise.

1.0.2 Discussion on Data Quality Issues and Measures

For AMG with respect to component-based modelling (i.e., the pragmatic use of data in this research), we assess data quality according to *how well data provides information for model component identification, composition and configuration.* Price and Shanks (2005) argue that the pragmatic use of data influences the perceptions of syntactics and semantic criteria. They includes this as a quality criterion in the pragmatic category. We support this view but do not explicitly include this criterion. We assume that the data use for AMG is a given goal, and data users shall consider the syntactics and semantic criteria with the application domain. Detecting and solving data quality issues need *domain understanding* and *data understanding*. The purpose of defining data quality categories and criteria is to help detect and solve data quality issues for AMG.

Data issues in the syntactic category are often straightforward. Syntactic accuracy (#1.) is related to the lawfulness rather than the correctness of data values (Wand and Wang, 1996). In many information systems, it can be automatically checked by *comparison functions* (Batini and Scannapieco, 2006). Syntactic consistency (#2.) is particularly relevant when data is sourced from multiple information systems (Shanks and Corbitt, 1999). Syntactic inconsistency can be typically solved through data type and format conversion.

In order to measure semantic accuracy (#3.) of a data value v, (i) the corresponding true value v' has to be known, or (ii) it should be possible, with the support of additional knowledge, to deduce whether v is or is not v' (Batini and Scannapieco, 2006). The first option is a non-option in a computational sense, because if the "true value" is or can be known digitally, then that value should be used instead of v. Hence, semantic accuracy is only computationally measurable and solvable with sufficient knowledge to reason the deduction.

Data completeness (#4. and #6.) issues can be found in both semantic and pragmatic categories. In either case, when data is truly incomplete[8], we can only complete the data by acquisition of the missing parts. Improving semantic completeness could potentially increase the chance of pragmatic completeness. Nonetheless, semantic incompleteness does not necessarily signify pragmatic incompleteness.

Mapping consistency (#5.) issues typically occur among data across different sources. Sometimes mapping consistency is broken because of erroneous schema changes (Velegrakis et al., 2004). When key values intended to map to the same external instance are inconsistent, a mapping table can be provided to clarify the relations among these keys.

Time can affect the validity of data. Given a time frame of data validity, timeliness (#7.) is easy to measure if data has metadata or dedicated fields (e.g., timestamps) to indicate its time attributes, e.g., when is the data collected or updated and how long is it valid.

Presentation suitability (#8.) particularly type-sufficiency poses many data issues in AMG. Type-sufficiency differs from semantic or pragmatic incompleteness in that the missing information can be deduced from the existing data with sufficient domain knowledge. The data is not (truly) incomplete but the information directly contained is not of the right type. When the domain knowledge and reasoning for deduction can be formalized, we are able to obtain the right type of information automatically from the data. This can be achieved through model transformation discussed in Section **??**.

To conclude, with regard to the information of model structure and parameterization, the data that is provided to the AMG as input should be assessed according to how well the data provides information for model component identification, composition and configuration. The requirements for the data (assume that the date has syntactic accuracy and timeliness) should have: (1) semantic and pragmatic completeness, and (2) syntactic and mapping consistency, or conversion rules or mapping tables or alike that can solve the inconsistency in the data. Transformation rules for AMG can be defined when modellers have sufficient domain knowledge and deductive reasoning to solve issues related to semantic accuracy and presentation suitability.

2 Conclusions

Acknowledgement

References

Ballou, D. P. and H. L. Pazer (1985). "Modeling Data and Process Quality in Multi-Input, Multi-Output Information Systems". In: *Management Science* 31.2, pp. 150 –162.

Batini, C. and M. Scannapieco (2006). *Data Quality: Concepts, Methodologies and Techniques.* Data-Centric Systems and Applications. Springer-Verlag Berlin Heidelberg.

[8] Meaning that (i) the data values or records are unknown but do exist, (ii) they are not contained by other accessible data sources, and (iii) they are not deducible from known data values or records.

Batini, C., C. Cappiello, C. Francalanci, and A. Maurino (2009). “Methodologies for data quality assessment and improvement”. In: *ACM Computing Surveys* 41.3, 16:1 –16:52.

Bovee, M., R. P. Srivastava, and B. Mak (2003). “A conceptual framework and belief-function approach to assessing overall information quality”. In: *International Journal of Intelligent Systems* 18.1, pp. 51 –74.

Catarci, T. and M. Scannapieco (2002). “Data Quality under the Computer Science Perspective”. In: *Archivi & Computer* 2.

Fox, C., A. Levitin, and T. Redman (1994). “The notion of data and its quality dimensions”. In: *Information Processing and Management* 30.1, pp. 9 –19.

Gelbstein, E. (2003). “Data, Information, and Knowledge”. In: *Encyclopedia of Information Systems*. Ed. by H. Bidgoli. New York: Elsevier, pp. 469 –476.

Huang, Y. (2013). “Automated Simulation Model Generation”. PhD thesis. Delft Univeristy of Technology.

Levitin, A. and T. Redman (1995). “Quality dimensions of a conceptual view”. In: *Information Processing and Management* 31.1, pp. 81 –88.

Loshin, D. (2011). *The Practitioner’s Guide to Data Quality Improvement*. Morgan Kaufmann.

McGilvray, D. (2008). *Executing Data Quality Projects: Ten Steps to Quality Data and Trusted Information*. Morgan Kaufmann.

Pipino, L. L., Y. W. Lee, and R. Y. Wang (2002). “Data quality assessment”. In: *Communications of the ACM* 45.4, pp. 211 –218.

Price, R. and G. Shanks (2005). “A semiotic information quality framework: development and comparative analysis”. In: *Journal of Information Technology* 20, pp. 88 –102.

Redman, T. (1996). *Data Quality for the Information Age*. Artech House.

Scannapieco, M., P. Missier, and C. Batini (2005). “Data Quality at a Glance”. In: *Datenbank-Spektrum* 14.

Shanks, G. and B. Corbitt (1999). “Understanding Data Quality: Social and Cultural Aspects”. In: *Proceeding of the 10th Australasian Conference on Information Systems*, pp. 785 –797.

Taylor, J. R. (1999). *An Introduction to Error Analysis: The Study of Uncertainties in Physical Measurements*. 2nd. University Science Books.

Velegrakis, Y., J. Miller, and L. Popa (2004). “Preserving mapping consistency under schema changes”. In: *The VLDB Journal* 13.3, pp. 274 –293.

Wand, Y. and R. Y. Wang (1996). “Anchoring data quality dimensions in ontological foundations”. In: *Communications of the ACM* 39.11, pp. 86 –95.

Wang, R. and D. Strong (1996). “Beyond Accuracy: What Data Quality Means to Data Consumers”. In: *Journal of Management Information Systems* 12.4, pp. 5 –34.