Data Quality—What Can an Ontological Analysis Contribute?

Andrew U. Frank

Department of Geoinformation Technical University Vienna Gusshausstrasse 27-29/E127-1 A-1040 Vienna, Austria

Abstract.
Progress in research on data quality is slow and relevance of results for practice is low. Can an ontological analysis make significant contributions? The "road block" in data quality research seems to be an ontological one. Approaching "data quality" with an ordinary language philosophy method reveals the inherent contradiction in the concept. The ontological analysis reveals the necessity to separate the ontology (reality) proper from the epistemology (data).
Data quality reveals itself when data is used, which focuses our attention on the double linkage between reality and data: (1) the observation that reflects reality into the data and (2) the decision that links the plan to the changes in reality.
The analysis of the processes leading from raw observations to decisions leads to operational definitions for "fitness for use" and an effective method to assess the fitness of data for a decision. Novel is the consideration of data quality as transformation through the whole process from data collection to decision.

Keywords: Ontology, fitness for use

1. Introduction

The importance of data for our society has increased immensely in the last 50 years. Data is a crucial element of economy, a resource comparable to physical resources like oil and steal. But unlike physical resources, for data no operational concept of data quality exists. Research progress is slow, despite efforts , e.g., a series of meetings originated by NCGIA [8; 9] or two biannual conference series ISSDQ [16; 17] and accuracy [1; 2].

Surveys have shown that relevance of the standardization that followed the research results is low [5].. Data quality descriptions represent the viewpoint of the data produces and are not very helpful for the potential data user to decide if a dataset is "fit for use".

A conceptual "road block" explains this unfortunate situation. The framework that organizes research is not adequate to capture the practical problems. Progress is not achieved by more efforts but by a refocusing after a fundamental, i.e., ontological analysis of the elements involved.

2. Approach

Applying the method of ordinary language philosophy [20; 10], we ask in section 3: “What does a person mean when saying 'this data is of high quality'?” This will reveal the inconsistency in the simple concept of data quality.

An ontological analysis in section 4 first points to the need to separate the reality, i.e., the ontology proper from the epistemology. The focus, in section 5, is then on the processes that connect reality realm and data realm: observations of the world produce data and decisions on actions use data to change reality.

With this conceptual framework section 6 shows how to use the insight practically in an assessment of the fitness for use of a dataset for a decision. Section 7 gives then a procedural, formalized, and programmable account of the assessment, followed by a final section with conclusions.

3. Ordinary Language Philosophy: What Does “High Quality” for Data Mean?

Ordinary language philosophy propagates as a method to start with ordinary, everyday speech and analyze its meaning [20; 10]. For the present question, one can start with utterances like “the road data is of high quality” and ask what the person making the utterance intends with it.

The meaning of the utterance depends on the situation:

- the person speaking produced the data and intends that the data were produced with great care and are—as far as possible—free of error and omissions.
- The person speaking uses the data to make a decision, for example in navigating with a car in a foreign city to a hotel, then it means that the data lead with little effort, and no uncertainty to the desired location.

The ordinary language approach clearly identifies the two conflicting interpretations of “data quality”: the viewpoint of the producer and the contrasting viewpoint of the consumer, already pointed out clearly by Timpf et al. [19].

A comprise seems to be reachable by interpreting the utterances in both cases as stating that the data is a true account of reality. This seems to capture the commonality between the two viewpoints. Equating “high data quality” with “true account of reality” leads unfortunately to another inconsistence:

Consider the level of detail included in a dataset. Obviously more detail makes a dataset more closely represent reality. Reality itself is infinitely detailed, for every data a more detailed dataset can be produced, which is of higher quality. The dataset with highest quality would be infinitely detailed and of infinite size. Several literary pieces have already pointed out the absurdity of a map scale 1:1 (e.g., Lewis Caroll, Jorges Luis Borge). The methods of ordinary language philosophy reveals inconsistencies in the term “high data quality”. The word “quality” is seemingly polysemous, with different meanings in different context.

4. Ontological Analysis

What are the fundamental elements that must be included in a meaningful account of “high data quality”? In the previous section the discourse includes “reality”, “data”, “production of data”, and “use of data in decisions” and the two persons performing these actions, these terms must be included at least in the analysis.

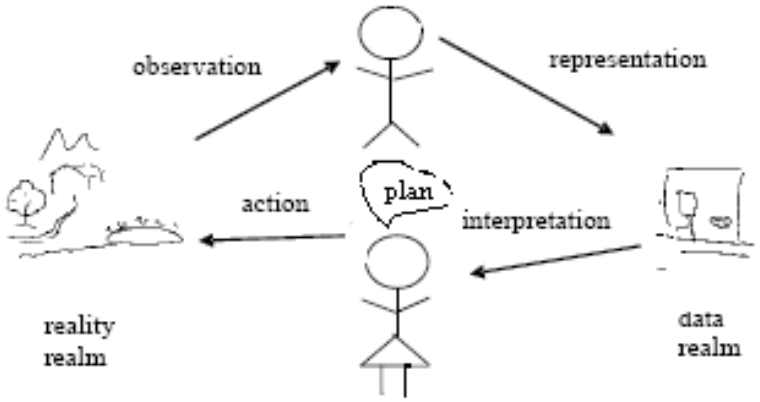

Figure 1: The reality realm and the data realm and the processes linking these

In the ontology necessary to discuss data quality—the *reality*, the object of ontological studies properly, and the *data realm*, the object of epistemological studies, is separated. This is not always done in ontological accounts, neither in philosophy nor in the technical use of ontology in the design of information systems [11]. Without the differentiation, the separation between reality and data describing reality, it is impossible to speak clearly about data quality, which qualifies the relations between the two realms.

5. Processes Link Reality and Data Realm

The focus must be on the processes which create links between the two realms. There is one process establishing the links between reality and data, ages creating data, and the other process going from data, representing planned states of reality to actions to realize these plans.

5.1. Observation

A tiered ontology [6] starts with an observable reality and posits that properties at points in space and time can be observed. These observations have a scale, obtained through the physical size of the observing system, and are influenced by random errors [7]. Example of point observations are remote sensing images, but also the human eye produces first point observations. Humans form out of point observations objects and integrate point properties to produce object attribute values. Objects and their attributes are typically constructed to be more permanent than the rapidly changing point observations. For example, in a taxi dispatch information system, color, form, etc. of a taxi cab are slowly changing, and only the location changes rapidly [12]—and the data is much more compact than the raster image of a movie of the taxi cab. The transformation of point observations to object attributes achieves an enormous data reduction; it is crucial for most human rational thinking.

Neither point observations nor object attribute data give a true (isomorphic) image of reality; the volume reduction when transforming point observations to object attribute values comes at a price, namely loss of

detail. The reduction to object data is guided by the situation and the intention of the observer, his experience, etc. The reduction achieved through the conceptualization of reality as objects is culturally influenced and multiple representations of reality as object descriptions are possible [7].

5.2. Decisions and Actions

Humans plan in the data realm desirable situations. Such plans are evaluated, decided upon and then appropriate actions to change reality to conform to the plan taken. This decision process is the (only) use of data. The decision process is limited in the sense of bounded rationality [18] and can be described summarily as: one or multiple future states (plans) are constructed mentally or in an external data representation; the plans are firstly checked for executability, i.e., are there physical or other laws impending the execution; executable plans are then secondly evaluated for costs and benefits and the optimal one selected for execution.

The data employed in such a decision process is subject to various errors that may result in an outcome of the actions quite different from the planned one.

5.3. Socially Constructed Data

Many of the facts important in todays economy and policies are socially constructed. Searle's formula “X counts as Y in context 2” [15] applies: a physical object or action, e.g., a round piece of metal, counts as a socially constructed object, e.g., as a one Euro coin, within a context, e.g., the part of Europe in which Euros are legal tender.

6. Is Data Fit for Use?

I start with the viewpoint of the consumer of data: data is of high quality if it is fir for use. This means in the framework outlined in section 5.2 that the decision process is leading to the desired result, specifically that the expected cost and benefits occur when executing the plan. The imperfections in the data can affect the decision in essentially two ways:

- executing the plan does not correspond to what was expected, in extreme, the plan cannot be executed;
- the plan executed is not the optimal plan and with better data, a different plan would have been selected.

The data producer must describe the data such that the consumer can determine whether data is fit for his use. The consumer needs to assess the effect the errors in the data have on his decision. The errors in the data can be classified in

- randomly distributed errors, with estimations of distributions applies mostly to observations an object attributes;
- effects of classifications, where fuzzy logic is [21] applicable (this applies mostly to object data);
- socially constructed data is often authoritative or guaranteed [4].

The idea of bounded rationality is to identify by how much a decision would improve if better data were available; if the effort to acquire this data is more than the improvement than it is not worthwhile to acquire better data. Of course, this assessment is again affected by risk and a cursory evaluation is sufficient (to avoid an infinite regress theoretically possible).

7. A Programmable Assessment for “Fit for Use”

The core of the assessment is the evaluation function used to make the decision. This function, especially the weights introduced to quantify to translate costs and benefits to the same scale is not always explicitly available, but Achatschitz [3] has shown methods to elucidate the evaluation function with preferences from the user.

All data element relevant for a decision for each data element and figure in the evaluation function. The error and risk associated with this data is introduced in the formula and then propagated to the evaluation to see how it influences this value. This is particularly easy for error with a random distribution: apply Gauß' error propagation law.

And similar risk propagation rules can be established for classification (fuzzy) data [7]. A detailed account how to propagate data quality to take influence on a decision can be formalized; the analysis shown that at least for engineering decisions, the arbitrariness of security margins eclipses all other uncertainties [14; 7]. Engineering decisions are reducable to a comparison load where the security factors f_c and f_e are fixed by conventions, standard, or laws. In the long run, a more risk oriented approach could increase security at lower cost [14].

As a rule of thumb, data is fit for use in a decision if the error and risk introduced by this dataset is less than $^1/_{10}$ of the total error and risk of the decision. For dataset where it appears difficult to assess the influences by formal methods of an error, one may ask questions to the human expert to quantify the possible influence or the increase in the risk of the decision.

8. Conclusion

It appears that data quality research is hindered today by some confusion in the notion “data quality”. An ontological analysis using methods from ordinary languages philosophy reveals that the term means different things in different context, even though there is a common element in all interpretations. An enumeration of necessary ontological categories to identify the meaning shows that it is necessary to identify two realms, the one of reality and the one of data, which are the two philosophical categories. The focus for data quality research is on the processes that link reality and data.

Data quality research has extensively studied how observations are made and data produced. Data producers do know the processes they use and the characteristics that serve to describe the quality of data from a data producers point of view. Not an equal effort is applied to understand how data is used. A model of a decision process, sufficiently simplified, based on bounded rationality leads to an operational method to assess the fitness for use of data:

A decision is based on an evaluation function that is used to determine the optimal choice. The imperfections in the data have effects on the decision when they change the outcome of this evaluation function sufficiently that another choice would be optimal or if a choice appears feasible when it cannot be realized in actuality.

For many datasets, types of data imperfection, and evaluation functions, the influence of imperfections on the outcome of the evaluation function can be formalized; for example, error propagation for randomly distributed error, fuzzy logic for classifications, supervaluation for radial categories [13]. When no data quality is available human experts may be asked to estimate the possible influence and risk. In general

humans have well developed senses for such effects, because most of our day by day decision making is done with incomplete and imperfect information.

It appears that the simple decision model included here provides the link between the data quality description of producers and the assessment of fitness for use by users. Novel in the approach to data quality is the “start to end” account, from data collection to decision made. This justifies the operational rule. The widely used rule by practitioners that influences which produce less than $^1/_{10}$ of the total uncertainty can be ignored.

9. References

[1] Accuracy (1991). *Proceedings of the Symposium on /Spatial Database Accuracy*, Melbourne, University of Melbourne.

[2] Accuracy. (2008). "Spatial Accuracy Meetings." from http://2008.spatial-accuracy.org/.

[3] Achatschitz, C. (2008). *Preference-Based Visual Interaction Spatial Decision Support in Tourist Information Systems*. Vienna, Technical University Vienna. Doctor.

[4] Bédard, Y. (1986). *A Study of the Nature of Data Using a Communication-Based Conceptual Framework of Land Information Systems.* orig.

[5] Boin, A. T. and G. J. Hunter (2007). What communicates quality to the spatial data consumer? *Proceedings 5th International Symposium on Spatial Data Quality*, Enschede, The Netherlands.

[6] Frank, A. U. (2001). "Tiers of Ontology and Consistency Constraints in Geographic Information Systems." *International Journal of Geographical Information Science* **75(**5 (Special Issue on Ontology of Geographic Information)): 667-678.

[7] Frank, A. U. (2008). "Analysis of Dependence of Decision Quality on Data Quality." *Journal of Geographical Systems* **10**(1): 71 - 88.

[8] Goodchild, M. and R. Jeansoulin, Eds. (1998). *Data Quality in Geographic Information - From Error to Uncertainty*. Paris, Hermes.

[9] Goodchild, M. F. and S. Gopal (1990). *The Accuracy of Spatial Databases*. London, Taylor & Francis.

[10] Grice, P. (1989). *Studies in the Way of Words.* Cambridge, Mass., Harvard University Press.

[11] Gruber, T. (2008). Ontology. *Encyclopedia of Database Systems*. L. Liu and M. T. Özsu, Springer-Verlag.

[12] Güting, R. H., Michael H. Böhlen, Martin Erwig, Christian S. Jensen, Nikos A. Lorentzos, Enrico Nardelli, Markus Schneider and J. R. R. Viqueira (2003). Spatio-temporal Models and Languages: An Approach Based on Data Types. *Spatio-Temporal Databases: The CHOROCHRONOS Approach*. Berlin, Springer: 117-176.

[13] Rosch, E. (1973). On the Internal Structure of Perceptual and Semantic Categories. *Cognitive Development and the Acquisition of Language*. T. E. Moore. New York, Academic Press.

[14] Schneider, J. (2000). "Safety - A Matter of Risk, Cost, and Consensus." *Structural Engineering International* **10**(4): 266-269.

[15] Searle, J. R. (1995). *The Construction of Social Reality*. New York, The Free Press.

[16] Shi, W., M. Goodchild and P. Fisher, Eds. (1999). *Proceedings of the International Syposium on Spatial Data Quality '99*. Hong Kong, Department of Land Surveying and Geo-Informatics.

[17] Shi, W., M. F. Goodchild and P. F. Fisher, Eds. (2003). *Proceedings of The 2nd International Symposium on Spatial Data Quality '03*. Hong Kong, Hong Kong Polytechnic University.

[18] Simon, H. (1956). "Rational Choice and the Structure of the Environment." *Psychological Review* **63**: 129-138.

[19] Timpf, S., M. Raubal and W. Kuhn (1996). Experiences with Metadata. *7th Int. Symposium on Spatial Data Handling, SDH'96*, Delft, The Netherlands (August 12-16, 1996), IGU.

[20] Wittgenstein, L. (1960). *Tractatus logico-philosophicus*. London, Routledge & Kegan Paul.

[21] Zadeh, L. A. (1974). "Fuzzy Logic and Its Application to Approximate Reasoning." *Information Processing*.