Data Quality Ontology: An Ontology for Imperfect Knowledge

Andrew U. Frank

Institute for Geoinformation and Cartography, Vienna University of Technology
Gusshausstrasse. 27-29, A-1040 Vienna, Austria
frank@geoinfo.tuwien.ac.at

1. **Abstract.** Data quality and ontology are two of the dominating research topics in GIS, influencing many others. Research so far investigated them in isolation. Ontology is concerned with perfect knowledge of the world and ignores so far imperfections in our knowledge. An ontology for imperfect knowledge leads to a consistent classification of imperfections of data (i.e., data quality), and a formalizable description of the influence of data quality on decisions. If we want to deal with data quality with ontological methods, then reality and the information model stored in the GIS must be represented in the same model. This allows to use closed loops sematics to define "fitness for use" as leading to correct, executable decisions. The approach covers knowledge of physical reality as well as personal (subjective) and social constructions. It lists systematically influences leading to imperfections in data in logical succession.

1 Introduction

Data quality, treatment of error in data and how they influence usability of data is an important research topic in GIS. It was research initiative I 1 of the NCGIA research plan {, 1989 #4636}. The lack of a formalized treatment of data quality limits automatic discovery of data sets and interoperability of GIS {Vckovski, 1997 #9564}. International efforts to standardize metadata to describe quality of data underline its practical relevance. Despite interesting specific research results, progress has been limited and few of the original goals set forth in the NCGIA program has been achieved.

Ontology is advocated in GI Science as a method to clarify the conceptual foundation of space and time. An ontology for space was the goal of NCGIA research initiative I 2 {NCGIA, 1989 #4636} and the later initiative I 8 extended this to time (to approach space and time separately was most likely a flaw in the research program). Ontology research in GIScience has led to classifications in the formalizations and representation of spatial and temporal data. Ontologists have only rarely considered the imperfections in our knowledge {Wand, 1996 #6693; O'Hara, 2001 #11015} and therefore not contributed to data quality research. Ontology claims to be useful to lead to consistent conceptualizations and to classify the design of GIS {Fonseca, 1999 #9640; Fonseca, 2002 #10084}. This paper demonstrates this contribution to the data quality discussion.

This paper is structured as follows: Section 2 reviews the ontological commitments of ordinary, perfect knowledge. Section 3 to 7 then detail in turn the ontological commitments with respect to practical causes for imperfection:

- Limitation to partial knowledge,
- observation (measurement) error,
- simplification in processing (object formation),
- classifications, and
- constructions.

Section 8 shows how these commitments for imperfection affect decisions. The application of ontological methods to data quality questions makes it necessary to combine the philosophical approach of an ontology of reality with the information science approach of ontology as a comceptualization of the world. It is necessary to consider both reality and the information model at the same level (an approach I have advocated for agent simulations before {Frank, 2000 #9607}.

Closed-loop semantics can then be used to ground semantics—including the semantics of data quality. This gives a novel definition of data qulity as "fitness for use" in the sense of leading to correct, executable definitions; this definition is operational and leads to method to assess the influence of imperfections in the data to errors in the decision {Frank, to appear 2007 #10876}.

2. An Ordinary GIS Ontology

The philosophically oriented ontology research debates different approaches to ontology. Smith has given several useful critiques of philosophical positions {Smith, 2004 #10963}. I use here the practical notion of ontology as a "conceptualization of a [ruler?] of reality for a purpose" {Gruber, 2005 #10822}. [quote not found]

2.1. Different Kinds of "Existence"

Ontology is concerned with what exists or is thought to exist. It appears obvious that the current temperature here, the weight of this apple, democracy in Austria and my loyalty to my employer each exist in a different way and follows different rules in their evolution {Medak, 1999 #9820; Medak, 1997 #8968}. A tiered ontology separates these forms (Fig. 1); I have suggested five tiers {Frank, 2001 #9957; Frank, 2003 #9920} and their influence on the conceptualizations and, as will be shown here, imperfections in the data will be discussed in this paper.

Tier O: human-independent reality
Tier 1: observation of physical world
Tier 2: objects with properties
Tier 3: social reality
Tier 4: subjective knowledge

Fig. 1. The five tiers of ontology

2.2. Ontological Commitments

Ontologies should be structured by commitments that state precisely assumptions used in the construction of the ontology. Consequences from the assumptions becomes transparent and contradictions between commitments can be avoided. Ontological commitments play in ontology the role of axioms in mathematics; from a different set of commitments different ontologies result (Smith and Grenon give a brief overview over different philosophical –ism resulting from different commitments {Smith, 2004 #10963}).

2.3. Commitments to a physical reality

2.3.1 Commitment O 1: A single world

It is assumed that there is a physical world, and that there is only one physical world. This is a first necessary commitment to speak meaningfully about the world and to represent some aspects in a GIS. This does not exclude description of future planned states of the world or the differences between the perception of individuals, because these are descriptions and not (yet) part of reality.

2.3.2 Commitment O 2: The world exists in space and evolves in time

The world exists in space and has states that change in time. This commitment posits both a continuous space and time. The current world states (unidirectionality of time); the first law of geography applies: all things influence all things, but nearby things influence more [Tobler ?]

2.3.3. Commitment O 3: Actors can observe the observable states of the world

The actors, which are part of the physical reality, can observe some of the states of the world. Observations give the state at a point in space and the current time (point observation)

$$v = p(\underline{x}, t) \quad \textbf{(1)}$$

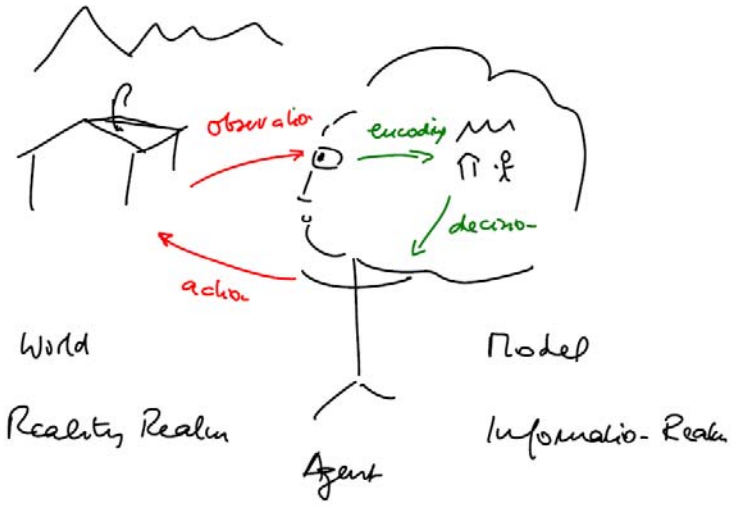

Fig. 2. Reality and Information Realm

2.3.4 Commitment O 4: Actors can influence the state of the world

Actors can not only observe the states of the world but also influence then through actions. The effects of actions are changed states of the world and these changed states can again be observed. This gives a closed semantic loop. [Frank 2003 a ?] which connects the meaning of observations through sensors with the meaning of actions that are reported through proprio-sensors (Fig. 2). The agent with the sensors and actuors in his body give semantic grounding to the observations and actions [Brochs – Body – Roboter].

2.4. Commitments Regarding Information

To construct a usable ontology for GIS, reality must be separated from the information system. Wang and Wand {, 1996 #6693} in perhaps the first ontological view on data quality have introduced this separation. This is in contrast to philosophical ontologies that attempt to explain reality and critique the concept-orientation of information system ontologies [Smith Fantilogy?], but also in contrast to Gruber's definition, which considers only the information realm.

2.4.1. Commitment I 1: Information systems are Models of reality

Observations and encoding translate the observable states of reality into symbol is the information system. The information system is constructed as a model of reality, connected by morphism, such that corresponding actions in model and reality have corresponding results. [Kuhn 91?, wand 91, Ceuster, Goguen 06]

2.4.2. Commitment I2: Information causation is different from physical causation

The changes in the state of the world are modeled by physical laws, e.g.: The cause for water flowing downward in the reality realm is gravity. The rules of physics can be modeled in the information realm and allows the construction of expected future states in the information realm, predicting what effects an action has.

An entirely different form of causation, is *information causation*. Agents use information to plan actions. The execution of the action causes changes in the physical world (O5). Actions can be separated into an information process, which I will call decision, and a physical action. Decisions are in the information realm but they affect—through actions and physical laws—the reality realm. Decisions can have the intended effect only if the action can be carried out and no physical laws contradict it.

2.5. Consequences

These six commitments are realistic, and correspond with our day-to-day experience. They imply other consequences, for example, O 4 with O 1 allows agents to communicate through actions (noise, signs), which are observed by others.

3. Ontological Commitments for Partial Knowledge

Perfect knowledge of the state of the world is not possible. The above "usual" ontological commitments ignore the necessary and non-avoidable imperfections in our knowledge. For many applications it is useful to pretend that we have perfect knowledge, but ignoring the imperfections in our knowledge is hindering the construction of a comprehensive theory of data quality and indirectly the realization of multi-purpose geographic information systems and interoperability between GIS.

In this section, three commitments regarding the incompleteness of our knowledge are introduced; they state limitations in

- spatial and temporal extent,
- type of observation, and
- level of detail.

3.1. Commitment P 1: Only a Part of the World is Known

Our maps do not show white areas as "terra incognita", but large part of the real world are still unknown (e.g., the microbiological realm). A complete and detailed model of the world would be at least as big as the world; it is therefore not possible. Our data collections are always restricted to an area of space and a period of time.

3.2. Commitment P 2: Not All States of the World are observed

A data collection is limited to the states observable with the sensors used; other states go unobserved and remain unknown.

3.3. Commitment P 3: The Level of Detail Is Fixed

Observation is a physical process and requires finite space and time for observation. The properties of the physical observation process limits the level of detail with which the (nearly) infinitely detailed reality can be observed. It is possible to obtain more detailed observations with better instruments, but not everywhere and all the time.

4. Imperfections in Observations

The observation methods achieve the first step in translating states of the world into symbols in the information system; they are imperfect physical systems and produce imperfect results.

4.1. Commitment E 1: Observations Are Affected by Errors

The physical observation processes are disturbed by non-avoidable effects and produce results that do not precisely and truly represent the state observed. These effects can be modeled as random disturbation [?] of the true reading and are assumed normally distributed with mean 0 and standard derivation σ.

4.2. Commitment E 2: States of the World Are Spatially and Temporally Anticorrelated

Nearly all world states are strongly anticorrelated, both in space and time. As the value of a state just a bit away or a bit later is most likely very similar to the value just observed. Spatial and temporal autocorrelation are crucial for us to make sense of the world and to act rationally, goal directed. Besides spatial and temporal autocorrelation, correlation between different observation types are also important. One can often replace a difficult to observe state by a strongly correlated, but easier to observe one; for example, color of fruit and sugar content are for many fruit correlated.

5. Imperfections Introduced by the Limitations of Biological Information Processing

Additional imperfections are the result of limitations of biological information processing and how it adapted to the environment in which biological agents survive.

5.1. Commitment R 1: Biological Agents Have Limited Information Processing Abilities

The structure of our information is not only influenced by our sensors, but also by the systems to process information. The brains of biological agents—including humans—are limited and the biological cost, i.e., energy, consumption of information processing, is high. Biological agents have therefore developed methods to reduce the load on their information processing systems to allow efficient decision making with limited effort and in short time. Humans and higher animal species reorganize the low level *point observations* into information about objects. Karminoff-Smith has shown that humans apply a general cognitive mechanism of re-representation of detailed data into more compact, but equally useful, information {Karmiloff-Smith, 1995 #10887; Karmiloff-Smith, 1994 #10898}.

5.2. Commitment R 2: Reduction of Input Streams by Concentration of Discontinuities

Most of the world is slowly and continuously changing (E 2); concentrating on discontinuities in this stable backdrop allows enormous reduction of processing load. This is the same strategy used in technical system (JPEG, run length encoding, etc.) and draws attention to discontinuities, i.e., boundaries, between larger relatively uniform regions. This is applicable both statistically for spatial regions of uniform color, texture, or movement, or dynamically, for uniform periods of light, cold, or economic well-being.

5.3. Commitment R 3: Object Centered Representation

Boundaries separate relatively uniform regions of some property from regions with some other value for the same property (Fig. 3).

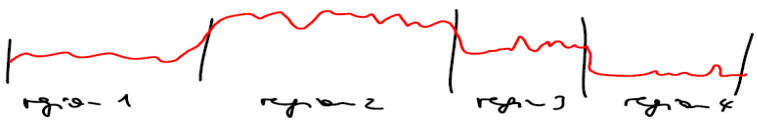

Fig. 3. Discontinuities separate uniform regions

The relative uniformity in the property value absorbs some of the observation uncertainty (E 1).

5.4 Commitment R 4: Objects Have Properties Derived from Observable Point Properties

It is possible to aggregate—usually sum—the values of some observable point property over the (2D or 3D) region of the object.

$$P(o) = \iiint_{V(o)} p(v)\, dV \tag{2}$$

5.5. Commitment R 5: Objects That Endure in Time Are Preferred

Objects that remain relatively unchanged over time are preferred to reduce the data processing[?] load. Material objects have slap boundaries where coherence in material is low: the object is what moves together. Such objects have a large number of stable properties (color, weight, shape, etc.).

5.6. Commitment R 6: Object Formation Is Influenced by Interaction

We cut the world in objects that are meaningful for our interactions with the world {McCarthy, 1969 #8061, 33}.Our experience in interacting with the world has taught us appropriate strategies to subdivide continuous reality into individual objects. The elements on the tabletop (Figure 5) are divided in objects at the boundaries where cohesion between cells is low and pieces can be moved individually; spoon, cup, saucer each can be picked up individually.

Fig. 4. Typical objects from tabletop space

5.7. Commitment R 7: Multiple Possibilities to form Objects

For a tabletop the boundaries of objects as revealed when moving them are most salient and it is difficult to form objects differently. At most different levels of aggregation are used: a pen with its cover is first one pen object, once opened, it is two (pen and cover) and occasionally one might take it apart and find more separatable preces [?]. For non-moveable objects no such salient boundaries are forcing a single, dominating form of object formation.

For example. Geographic space can be divided into uniform regions (i.e., geographic objects) in many different ways: uniform land-use, land-cover, watersheds, soil quality, etc. A GIS must be prepared to deal with multiple, overlapping, and coexistent geographic objects and must not blindly transfer the exclusive, non-overlapping solid object concept from tabletop space to geographic space [Montello COSIT Semmering?].

5.8. Commitment R 8: Error Formation Can Be Relative or Absolute

Object boundaries can either be induced by a property passing a threshold (Figure 6):

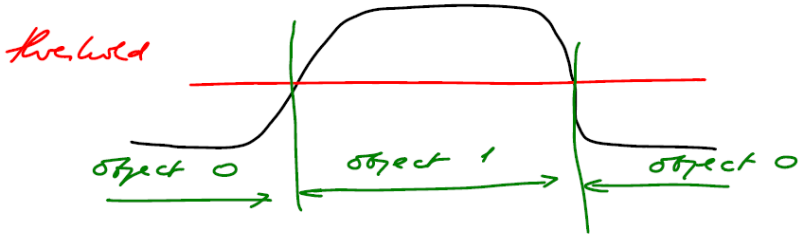

Fig. 5.

Or the boundary is where the change (gradient) is maximal (Figure 7).

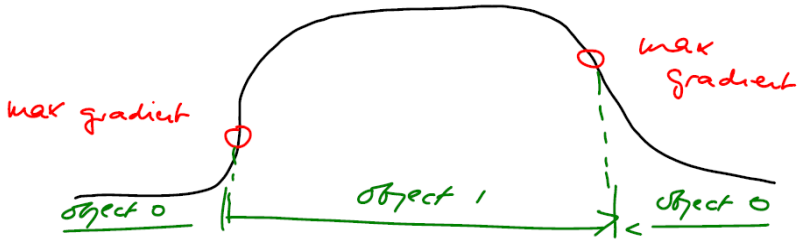

Fig. 6.

Technical and legal systems often use the fixed threshold (e.g.; mountain farming support is available for farms above 1000 m above sea level). Human sensors seem to identify changes independently of absolute value {Burrough, 1996 #254}.

5.9. Commitment R 9: Object Boundary Location is Uncertain

The measurement error (E 1) influence the position of the object boundary. Spatial (or temporal) autocorrelation is necessary to assess the effect.

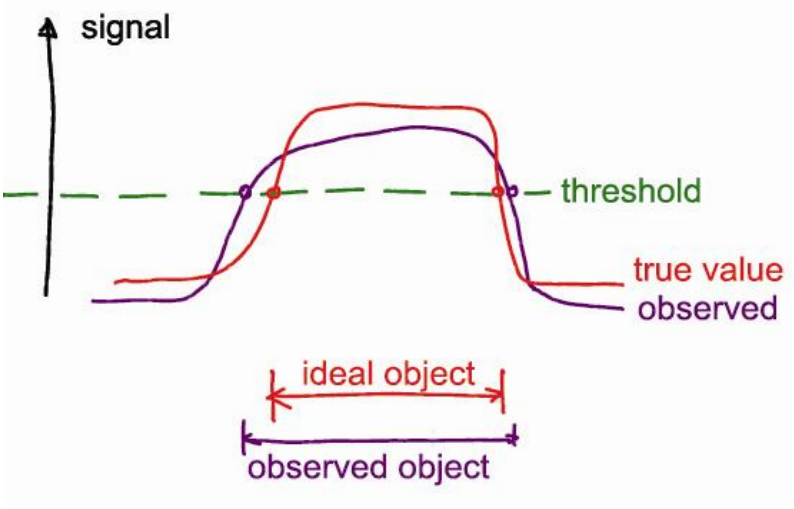

Fig. 7. Error in observation of property results in error of object boundary

5.10.Commitment R 10: Object Property Values Are Uncertain

The property value obtained for an object is an approximation influenced firstly by the uncertainty in the object region delimited by the object boundaries (R 9) and

secondly by the measurement error of the point property that is integrated (E 1, with Eq 2 from R 4).

6. Classification

The interactions effects have with the world seem to repeat themselves with very small variation in the particular objects involved: I get up every morning, drink tea and eat a toast—it is a different day, tea , and toast every morning, but I can regulate my behavior by using the same patterns every morning. The discussion is in terms of physical objects, but the mechanisms are used in the same way for processes.

6.1. Commitment C 1: Objects Are Classified

Physical objects, which result from commitments R 1 to R 9, are classified. Classification selects a property to determine object boundaries (e.g.; material coherence) and one or a set of Properties of the object to classify it {Frank, 2006 #10824}. Prototypical is the classification of animals and plants to species and the hierarchical sub/super class structure created by Linné.

6.2. Commitment C 2: Class membership of an Object is Based on Object properties

Two mechanisms are used for classification; they can be called set-theoretic (Aristotelean) and radical categories. A set-theoretic category includes all objects that fulfill some property (or properties), e.g., a tree is an individual plant from a tree species with a diameter of more than 10 cm at 1.5 m height above ground. All members of the class tree so defined have this property and no member does not.

$$\{x \in \text{tree} \mid \text{diameter } 1.5\,(x) > 0.10\} \quad (\mathbf{3})$$

A radial category is formed by a number of properties (among then often gestalt) and allows better, more typical and atypical members. Typical *birds* are robin or sparrow, penguin and ostrich are atypical.

6.3. Commitment C 3: Object Reasoning Can Use Default Class Values for Properties

In many situations detailed knowledge about the individual object is not available and replaced by the expected (default) value for the class. This permits reasoning in the absence of detailed knowledge, for example to analyze future situations, for which detailed knowledge is not available yet. Set-theoretic class are better suited for default reasoning; using default reasoning with a radial category may be wrong.

Tweady is a bird
Birds fly

$\Rightarrow$ Tweady flies

Is correct unless Tweady is an atypical bird (e.g., a hen).

6. Constructions

Searle has described social constructions with the formula: X counts as Y in context Z. This formula applies equally for personal (subjective), linguistic and social construction {Searle, 1995 #8011}.

7.1. Commitment X 1: No Freestanding Y Terms

All constructions are eventually grounded in physical object or process. It is possible that a construction 'B counts as A in context Z' and B is a construction and not physically existing but then B is part of another construction 'C counts as B in context Z', where C is a physical object or action. Any construction is ultimately grounded in a physical object or action.

7.2. Commitment X 2: A Context is a Set of Rules

The meaning of a Y term is determined by the context Z; the context Z describes a set of logical rules that give antecedents, consequences potential actions, etc., which are related to Y. Such a context can be modeled as an algebraic structure; Bittner and Navratil have given a formalization of the rules that form the context of landed property ownership and cadastral system {Bittner, 2001 #9848; Navratil, 2002 #10254}. The difference between personal (subjective), linguistic , and social constructions seem to be in the applicable context:

- A personal construction is valid in the context of a person, her experience, history, current situation and goals, etc. It is highly variable and not shared.
- A linguistic construction, a word in a language for example, is valid in the current cultural use of this construction; the context gives the rules for entailments; i.e., what another person may conclude from an utterance.
- A social construction in Searle's sense is valid in a shared, often codified set of rules; a national legal system is a prototypical example for such a context; it can be very complex.

8. Quality of Data

8.1. Quality of Observations and Object Properties

Observations can be assessed by repetition and statistical descriptions of errors and correlations derived (E 1 and E 2). If the rules for object formation are known, the quality of the data describing object properties can be described statistically as well (R 9 and R 10). Such descriptions represent a data acquisition point of view and are not directly relevant to determine if a dataset is useful for decision {Timpf, 1997 #6721; Timpf, 1996 #6717}.

8.2. Data Quality Influencing Decision

The important question is whether a dataset can be used in a decision situation, which can be restated as 'How do the imperfections in the data affect the decision?' For decisions about physical actions based on physical properties, as they are typical for engineers, statistical error propagation can be used.

$$\sigma_r^2 = \sigma_u^2\left(\frac{df}{du}\right)^2 + \sigma_v^2\left(\frac{df}{dv}\right)^2 + \sigma_w^2\left(\frac{df}{dw}\right)^2 \quad (4)$$

For example, a decision about the dimensions of a load bearing bean reduce to a comparison of resistance R with the expected load L:

$$R > L \text{ or } R - L > O \quad (5)$$

In general decision seems either to be of the 'go/no go' type, which reduce to a statistical test for a quantity more than zero (as shown above) or to a selection of the optimal choice from a number of variants. In practical cases the valuations involved are often unknown, but it seems possible to reconstruct the structure of the decision process after a decision has been taken and then to analyze the influence of data quality on the decision. Decisions usually include also default reasoning with categorized data; a reconstruction of the decision process as a formalization and the identification of the imperfections and their causes through the ontological commitments show what tools can be used: statistics with other than normal distributions, fuzzy set theory, assessment of differences in contexts etc.

9. Conclusion

An ontological view of data quality is different from a philosophical ontology—which concentrates on the existing world—but also different from the information science ontology—which concentrates on the concepts representing the existing world {Gruber, 2005 #10822}. It must include both the reality realm and the information

realm. Only if both realms can be linked conceptually in the closed semantic loop (Fig. 2) of observation and action, a meaningful discussion of data quality is possible.

Experience with decision processes has taught us the appropriate levels of data quality necessary to make the decision most of the time correctly. In a world of interoperable information systems connected by the web, an analytical and formalizable treatment of data quality is required.

[MOVE?]Assessment

Data about reality is collected to be used in a decision process. Experience has taught us what region and period is relevant, which observation and what level of detail are necessary to make optimal decisions.

10. References