ONTOLOGIES FOR IMPERFECT DATA IN GIS

Note: figures will be improved for publication

Andrew U. Frank
Institute for Geoinformation and Cartography
Technical University Vienna, Austria
Gusshausstrasse 27-29/127, A-1040 Vienna

Abstract

The importance for ontological clarification to design GIS, to structure data in a GIS or to construct usable user interface is well established; ontologies are crucial to extend interoperability from a syntactic to a semantic dimension. The discussion of ontology for GIS always pretends that the data represent reality perfectly, but real data in a GIS can give only an imperfect image of reality. An ontology for imperfect data is necessary, which is an ontology of imperfections in the representation. The analysis starts with a brief review of the ontology typically assumed for a GIS, followed by the description of the ontology of the unavoidable imperfections in the data collected. This covers aspects like partial knowledge, measurement errors, object formation, etc. (restricted to information about physical objects, e.g., data in a GIS with environmental purposes). An ontology of imperfections sheds new light on the quality of information discussion and leads to an operational definition for data quality not based on perfection. Sufficient quality of data is achieved if further improvements would not improve a decision noticeable. This leads to a differentiation of how insufficient data quality can influence a decision.

1 Introduction

Ontologies are necessary to clarify semantic aspects of geographic data (Frank 1983). Ontologies for GIS play a dominant role in the discussion to extend GIS interoperability (OGC 2000) from a syntactic to a semantic dimension (Kuhn Draft 2007) and numerous papers have been published on semantic or ontological alignment between two geographic data collections in order to achieve interoperability and to prepare for combining the data for answering a query. Ontologies are useful for communication with users in the design of a GIS (Fonseca et al. 1999) and to fix the data structure for the spatial database. Ontologies contribute to bridge the gap between user and GIS in the user interface (Kuhn 1993), and could contribute to make data quality information more usable (Comber et al. 2006; Boin et al. 2007b).

Ontological studies, both in the philosophical tradition and in the computer sciences, application of ontologies assume that the data collected give a perfect image of reality. The rare exception is a paper by Wand and Wang (1996), which suggests that data quality should be anchored ontologically. This paper is regularly cited in the information management literature, but I could not detect that it has influenced the philosophical computer science or geographic information ontology discussion.

Data in a GIS is never perfect and understanding and communicating the level of imperfection in data is key to interoperable use of data. In this article an ontology for imperfect data is systematically constructed. It uses as a foundation a realistic ontology, structured in tiers (Frank 2001; Frank 2003) and then assesses for each process of data acquisition and transformation between the tiers how it degrades a perfect representation to produce the imperfect data we have to deal with.

Data quality is recognized as crucial for interoperability and efforts to standardize metadata, including data quality information, led to international standards (DCMI 2006). These efforts are not using ontological approaches to clarify the semantics of the data quality dimensions, e.g., accuracy, timelines, etc. The ontological approach gives a much finer classification, which is operational because the definitions relate to the operations of data acquisition and processing. For example, most data quality description proposes a distinction between spatial and temporal error; such a distinction is not operational for moving objects, because temporal error can masquerade as error and vice versa. The imperfections in data are ontologically differentiated in

- incompleteness of our partial knowledge,
- definitions of observation process producing errors, and
- restrictions in cognitive processing forces on object-oriented approach,

each with several subclasses. This paper improves a first description of ontological commitments for imperfect information (Frank to appear 2007) and extends it. For lack of space, the treatment here does not yet address the important and very difficult problem of errors in classification, which is related to linguistic issues like the so-called prototype effect (Rosch 1973), the question of ‘language of thought’ (Fordor 1975; Jackendoff et al. 2002), etc. This article gives the foundation on which these questions can be addressed later systematically. A suggestion how I plan to extend the material exposed here was recently presented (Frank 2006).

From the shift of focus from errors in data to imperfections as data follows a new, operational definition of data quality. Data is used to make decisions and imperfect data lead to imperfect decisions—at least conceptually. Relating data quality to decision quality leads to a view that data is of sufficient quality for a decision if improvement of the quality of the data would not noticeably improve the decision, At the end of the paper a very brief example shows how the ontologically differentiated aspects of data imperfection affect a decision. This is only a preview for a systematic treatment of the influence of data quality on the quality of decisions; a preliminary conference contribution is published (Frank 2007).

This paper is structured as follows: Section 2 gives an overview on how ontology can contribute to achieve a logically consistent view of data quality and summarizes the usual assumption for a GIS assuming perfect knowledge. Section 3 then describes the commitments for an ontology for imperfect knowledge separating

- the necessary partial knowledge of an infinitely complex world,
- the unavoidable observation errors, and
- the strategies humans use to reduce the amount of data to process.

Section 4 compares the traditional definition for data quality, based on correspondence between reality and data, repeatability of observations, quality of the data acquisition process or ‘fitness for use’. Section 5 puts theses definitions in the context of decision making and leads to the investigation of the influence of data quality on a decision process in section 6. Section 7 gives a small example before coming to conclusions in the last section.

2 Ontology Contributes to Understanding Data Quality

Ontologies for information systems describe conceptualizations of a subset of reality for a purpose; they describe the conceptualization of the world used in a context (Gruber 1993). A context is formed by, for example, a Geographic Information System and the application that uses the data in the GIS. It is argued that ontologies contribute to the design of information systems, especially GIS (Fonseca et al. 2002). To achieve a consistent description, ontologists fix initially *ontological commitments*. Details of these commitments are extensively debated in philosophy; for an overview of the different “–isms” applicable to GIS see an article by Smith and Grenon (2004). Ordinary ontology does not take into account the imperfections in our knowledge of the world.

The clarification of ontological commitments leads to consistency in a geographic information systems; the same methods must be useful to clarify the quality of the data in the GIS. In this section I review the usual ontological commitments for a GIS, as a backdrop to discuss in the following section an ontology for data imperfections. I use here the tiered ontology to separate different kinds of physical or socially constructed existence in tiers (Frank 2001; Frank 2003) and restrict the discussion to the ontology of objects in the physical world, excluding social, cultural, and subjective constructions (Searle 1995).

2.1 Tiered Ontology

In the tiered ontology (Frank 2001; Frank 2003), five ontological tiers are separated (Figure 1): From these this article concentrates on the first three: (i) the reality, (ii) the observation of reality and (iii) the physical objects formed by cognitive agents. The treatment is restricted to physical aspects of reality and does not cover the subjective and social constructions that are important to organize the world. This limitation excludes difficulties introduced through the social conventions and contexts, which must be discussed later. The concentration here makes it possible to demonstrate some important concepts clearly before embarking on a discussion of data quality and its effects in social data. Aspects of these difficulties were discussed at a workshop and documented in an edited volume (Burrough et al. 1996). Gangemini et al. have published formal methods in the DOLCE framework in which social and subjective realities can be modeled (Masolo et al. 2003); special cases for mapping and for cadastre have been addressed by Frank (2000a) and Bittner (2001). These approaches must be extended with the same considerations for error, uncertainty etc. detailed in the next sections.

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

Figure 1: The five ontological tiers (Frank 2001)

2.2 Ontological Commitments to Conceptualize the World

This subsection reviews ontological commitments as they seem to be customary for physical geography oriented GIS.

2.2.1 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 multiple description of future planned states of the world or the differences between the perception of individuals.

2.2.2 Commitment O 2: The world exists in time and has evolving states

The world has states, which evolve in time (Wand et al. 1996). This ontological commitment is twofold: it posits a single time and changeable states of the world.

2.2.3 Commitment O 3: Observable states

The actors in the world can observe some of the states of the world at a given location and the present time (the *now* of Franck (2004)) with their sensors. Observation of physical state for properties of the world states at a point in space and time is objectively possible (*point observations*); the influence of the observer on the observation of physical properties is small and repeated observations give the same values.

2.2.4 Commitment O 4: Information systems are models of reality

Observation result in information and the ontology must separate the *reality realm* from the *information realm* (postulates 2 and 3 in Wang and Wand (1996, 90)). Observations translate the state of the world from the realm of reality to the realm of information (Figure 2). The information realm is a partial and incomplete *model* of the world as the commitments to imperfect knowledge in the next section postulate. A model is related to the world by morphism. Corresponding operations in the model have corresponding results (Kuhn et al. 1991; Wand et al. 1996; Ceusters et al. 2006; Goguen et al. 2006)).

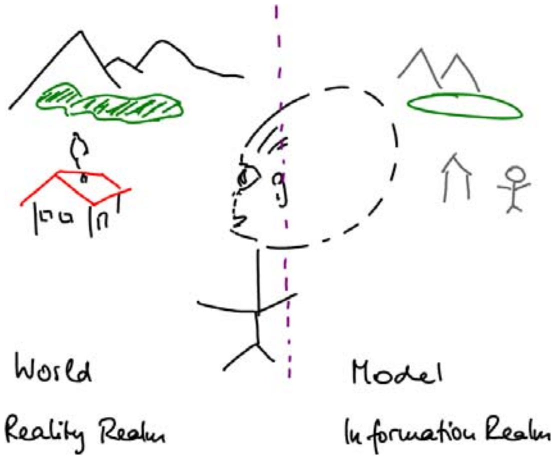

Figure 2: The Reality Realm and the Information Realm

2.2.5 Commitment O 5: Some states of the world can be changed by agents

The actors in the world can not only observe the world, but they can change it through actions. The effects of actions are changed states of the world and these changed states can be observed. This gives the *semantic loop* (Figure 3) that connects the sensors and their observations to the changes caused by the actions (Frank 2003). The semantic loop links the proprio-sensoric experience of actions with the sensors in the human brain.

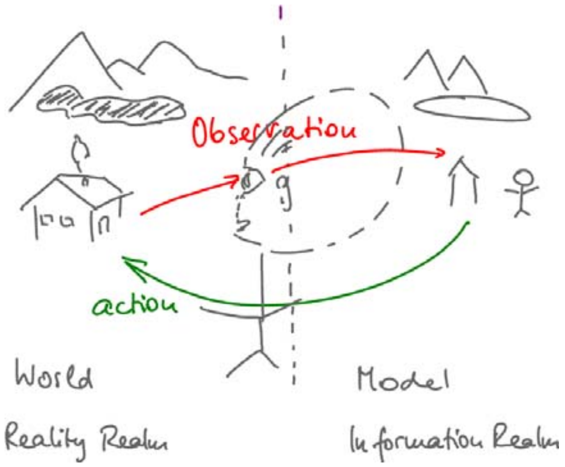

Figure 3: Closed Loop Semantics Connects the Reality Realm with the Information Realm through Observations and Actions

2.2.6 Commitment O 6: Separate physical and information 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 will have.

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 impede it.

3 Ontology for Imperfect Information

The above "usual" ontological commitments ignore the necessary and non-avoidable imperfections in our knowledge of the world. They pretend that we have perfect knowledge, which is not possible. The illusion of perfect information is acceptable for ordinary, every day operations, given that experience helps us collect sufficient information to make valid decisions, but ignoring the imperfection of our knowledge is hindering the construction of advanced information systems.

The limitations in the knowledge we have about reality can be divided in

- partial knowledge,
- impossibility to obtain true values,
- reduction of detail by formation of object, and
- classification.

This article concentrates on physical models of reality and concentrates on the first three influences on the acquisition of information and ignores effects of classification, which should be discussed in the context of social constructions (for a preliminary discussion see (Frank to appear 2007)). The ontology structured in tiers advances the analysis of imperfections in data. The imperfections are introduced by the processes that connect the tiers: first, the observation processes linking tier O to tier 1 and then object formation processes linking point observations in tier 1 to object descriptions in tier 2.

3.1 Partial knowledge

It is impossible to construct a complete model of the (nearly) infinitely detailed reality and our knowledge is therefore always a partial selection of those aspects that seem important. Three ontological commitments for partial knowledge P1 through 3 describe limitations which reflect the incompleteness of our knowledge:

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

3.1.1 Commitment P 1: Only information about a part of the world is collected.

Most geographic data collections focus on a small, spatially delimited area of the globe. We speak of the *geographic footprint* of a GIS and mean the area covered.

The same limitation applies in the time domain: a dataset gives the state of the world at a given time, usually the *time of data collection*. The time of data collection is well defined for remote sensing data, but much less clear for other datasets in GIS, which are constantly updated. The data collection time for different data elements varies and the “update level” is often differentiated by importance of a change (e.g., in the Bavarian ATKIS data collection changes in important roads are updated within 3 months, whereas forest roads lag behind for years (Henninger 2007)).

3.1.2 Commitment P 2: Not all properties of the world are observed

Assuming that the world has observable states, a single data collection contains only a subset of these. For example, remote sensing images observe the intensity of the light reflected in a set of frequency intervals. A collection of datasets is described by the *observations* they contain. The models we construct are restricted to the aspects that are relevant for the decisions we intend to make and we limit the expensive collection of data to aspects we deem relevant.

3.1.3 Commitment P 3: Information is limited by the level of detail observed

The world is infinitely complex and the information we have about it is always limited by the level of detail. It is impossible to construct a fully accurate and detailed model of the world, because such a model would be at least as big as the world. The information model of the world is therefore always limited by the *level of detail*.

3.2 Observations Do Not Produce True Values

The observation methods that translate states of reality to information are by themselves limited. The data collected are in error, which is described by the error commitments E 1 and E 2. Fortunately most changes in the world occur gradually—both in space and time—and therefore the human approach to partial observations with imperfect sensors yields usually valid information. Figure 4 shows the small effect an error in location has on the value of the signal, because values nearby are very similar. Human rational behavior would not be possible in an uncorrelated world! Strong correlation leads sometimes to confusion, e.g., land use and land cover are strongly correlated and often confused in data collection and classification.

3.2.1 Commitment E 1: Observations are erroneous

Observations are affected by unavoidable effects, which create differences between the ideal observation (the ideal value) and the actual realization of the observation. These effects appear as random and are modeled, unless one has better knowledge, by normal distributions with an expected value *v* and a standard distribution *σ*.

3.2.2 Commitment E 2: Autocorrelation

Autocorrelation counteracts the effects of the fact that all our knowledge is incomplete and erroneous. The physical world is strongly autocorrelated—both in space and time. This allows to interpolate values for points not observed or compensate observation errors. The most likely observation of a property just a little bit to the left where I looked before, or just a little bit later is most likely very similar to the observation I made before. Different observations are also strongly correlated: for example sugar content of a fruit and its color is usually correlated—we pick the red strawberries, because they are sweet and taste good and leave the green ones.

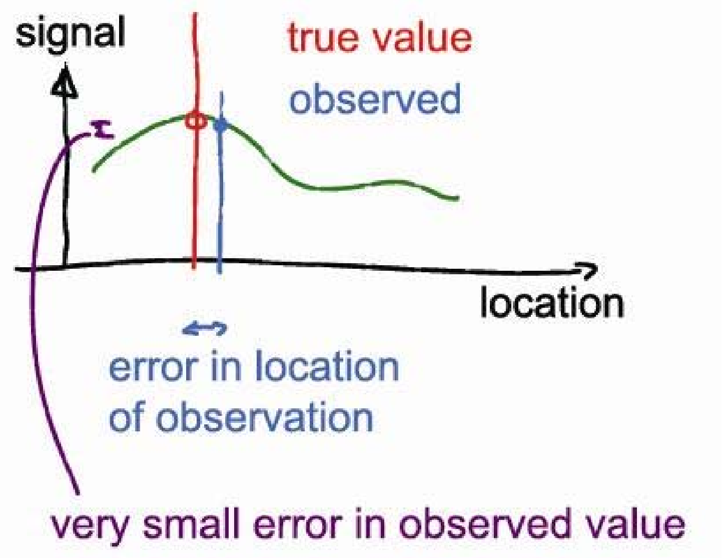

Figure 4: Large error in location leads to small error in observed value

3.3 Limitations of Information Processing: Forming Objects

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 1994; Karmiloff-Smith 1995). These limitations are captured in the restriction commitments R 1 through R 8.

3.3.1 Commitment R 1: Biological agents have limited information processing abilities, therefore concentrate on discontinuities

Most of the world is slowly and continuously changing and most aspects of the world are highly correlated (E 2). Focusing on the discontinuities gives a compact representation. Before a background of stable states of the world we focus on the interesting changing and discontinuous points. This re-representation reduces the data that needs to be processed and still gives sufficient information about the world to avoid dangers (e.g., to detect the rapidly approaching car). In technical systems autocorrelation is used to reduce bandwidth necessary for transmission, e.g., of television images; strong autocorrelation is the reason lossy compression methods like JPEG and MPEG or loss-less methods like run-length encoding work.

3.3.2 Commitment R 2: Object centered data processing

Human processing of information describing reality is object oriented. Humans, and many other biological agents, convert the visual observations they perceive in a raster format into information about objects. In GIS raster data is often converted into vector based object data. The observation of properties of points in space and time are restructured, i.e., re-represented, to become properties of objects.

The cognitive system forms objects at discontinuities (R 1): it is simpler to keep track of objects with uniform properties and to pay attention to their boundaries; most of the world modeled as objects is stable, uniform, and unchanging compared to a point (raster) model of the world. Due to high spatial and temporal autocorrelation large areas have similar properties and are thus reduced to a single object. This reduces the cognitive load.

3.3.3 Commitment R3: Object formation is driven by interaction with the world

We cut the world in objects that are meaningful for our interactions with the world (McCarthy et al. 1969, 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 and moved individually.

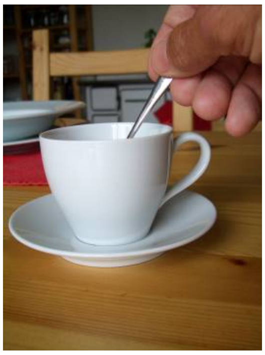

Figure 5: Typical objects from tabletop space

3.3.4 Commitment R4: Objects are formed to endure in time

The world is not static (O2), but changes are rare compared to the overwhelming relative stability. Object formation further increases this stability by separating seldom changing properties from frequently changing ones. For example, weight or color of the cup in Figure 5 is seldom changing, but location of the cup changes often.

3.3.5 Commitment R 5: Multiple ways to form objects

There is little doubt how to limit familiar movable objects, but variations are possible: in Figure 5, we can separate cup and saucer, but often the two are sold as one “coffee-cup” object; in a restaurant a cup of coffee includes also the coffee, cup, and spoon—they are served and billed as a unit. Even more variable are non-moving, geographic objects; there are multiple ways to subdivide geographic space into objects. In the absence of movement, a single, most salient criterion for object formation is not available (Frank 2006). Considering terrain, we can identify watersheds, and valleys, or we identify alpine meadows and forests, agricultural and built-up areas (Figure 6). These multiple ways to define objects do not necessarily coincide but often do, caused, e.g., by correlation between slope and land use. Many other ways to subdivide space are important; if we consider social constructions: ethnical and religious boundaries are seldom coincident, leading to wars about what is the natural limit of a territory. GIS must be prepared to have coexistent, overlapping spatial objects (Frank 2001).

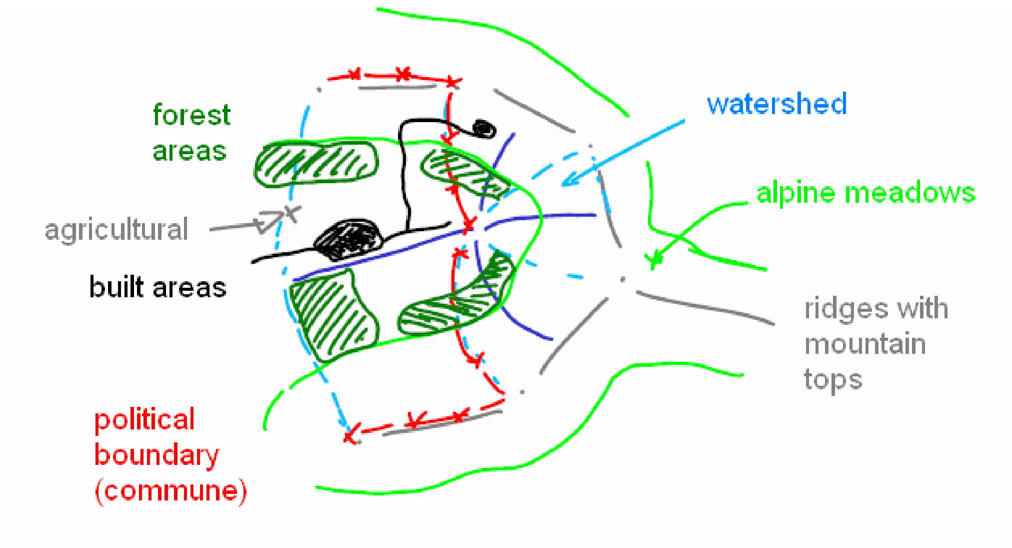

Figure 6: A region with watershed and land cover objects

3.3.6 Commitment R 6: Objects as regions with uniform properties

A very general approach to define objects is to say that they form regions with at least one attribute having a uniform value. For movable objects the uniformity is in the joint movement; this usually coincides with uniformity in material, color, etc.

The uniformity of the attribute allows for variation of the value of the property within limits and introduces *thresholds* for which the property value is considered “uniform”. This absorbs error in the observations (E 1) and reduces the necessity for precise observations.

Different objects result if we select different attributes. Areas of uniform land cover (e.g., alpine meadows, forest) do not necessarily coincide with the areas of uniform water flow that gives watersheds or regions of similar soil type (Figure 6). The autocorrelation in space and the correlation between factors influencing natural processes result in object boundaries that often (nearly) coincide. The differences in land cover coincide with the fences in Figure 7 and the differences in land use; the road along the pond nearly coincides with the boundary of water covered area.

Figure 7: Coincidence with land use changes

3.3.7 Commitment R 7: Object formation is uncertain

Objects, formed as areas of uniform attribute value, are delimited by boundaries and these boundaries have observational error. The error in observing the attribute affects the determination of the boundary (Figure 8).

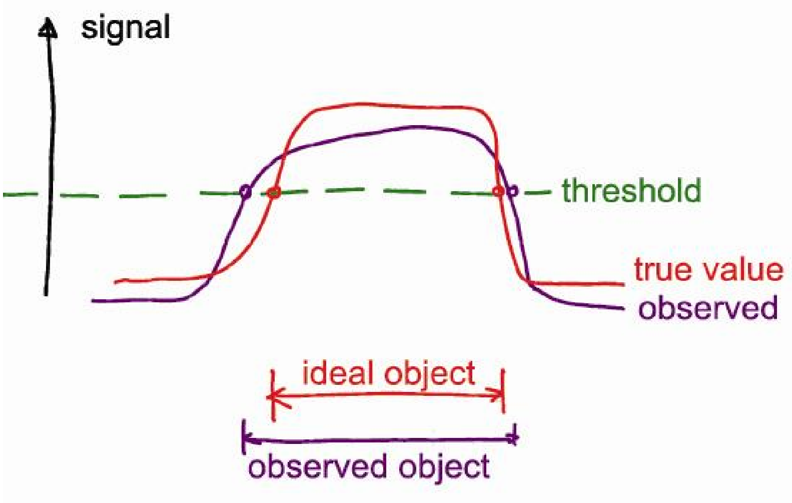

Figure 8: Error in observation of property results in error of object boundary

3.3.8 Commitment R 8: Objects have properties, known only with uncertainty

Objects have changeable state. The changeable state of objects is the consequence of the assumption that the world has changeable states (commitment R 2 and 3) and objects are formed as aggregates of real world regions (R6). The object properties summarize the observable point properties. Object properties are integrals of some attribute over the volume the object forms (Eq. 1) and describe the state of objects.

$$P(o) = \iiint\limits_{V(O)} p(v)\, dV \qquad \text{(Eq1)}$$

Object property values, as obtained through equation 1 are *approximations*. For small moveable objects, practical methods to observe these integrals directly exist. For example, weighting an object on a balance gives directly the integral of the mass over the object volume. Such methods are generally not available for geographic objects. The error in the observation of properties of the objects is affected by observational errors in two forms:

a) error in the determination of the object boundary (R 7) and
b) error in the observation of the attribute summed (O 5).

These errors can be modeled if the observation errors are assumed to be random, normally distributed (commitment E 1) (Navratil et al. 2006). However, simplifying assumptions are necessary to achieve a tractable formalization; for example, weak correlations between properties (E 2) are ignored.

4 Quality of Information

Given these imperfections, how to define quality of information? Obviously, information is of higher quality if the imperfections are lower, i.e., the combined influences of incompleteness, error, object formation, etc. are less; we have also seen effects that reduce imperfections, e.g., autocorrelation (E 2). The information with less imperfections, corresponds better with reality. Questions arise: How much quality is required? Do we always need perfect information? Obviously not—but how much is necessary? What does it mean precisely to say that information corresponds with reality? Four approaches to define data quality are found in the literature. Data quality is defined as:

- correspondences between reality and data,
- agreement between repeated observation,
- result of the quality of the data acquisition process, or
- ‘fitness for use’ of the data.

All four definitions of data quality have some difficulties and do not lead to operational, practically useful methods to assess data quality. These four definitions will be discussed in the following subsections. The next section then brings an integration of the four approaches and the following section sketches how this integrated concept oriented towards decision making can be made operational to be practically useful. The application to a more complex engineering decision is reported else where (Frank 2007).

4.1 Information Assumed Isomorphic to Reality

A common naïve but very useful assumption in the construction of information systems is that the information is an isomorphic image of reality, meaning that there is a one-to-one relation between reality and the data. This is essentially Tarski's correspondence theory for truth (Tarski 1944). If the two realms in Figure 2 are linked by an isomorphism, then the distinction between a thing in the world and an image in the information model would not be required. Reality and the model are the same—up to isomorphism (Lawvere 1965; Mac Lane 1998). This simplifies reasoning and talking about data enormously but does not produce a method to assess data quality. If we consider imperfection in the data, the simplification resulting from the assumption of an isomorphism is not justified and the mapping between reality and information model must be analyzed and not glossed over (Kent 1979). Over time, methods emerge to produce the quality required, but in today's climate of rapid development of technology and business practice, an analytical treatment of data quality is necessary to actively design methods to achieve the quality required and not more.

4.2 Quality Derived from Repeated Observations

The most often used definition of information quality is based on the *repeatability of observations*: another observer bringing back the same information demonstrates that it is correct. This "internal precision" leads to a quantitative assessment of data quality, used extensively in, e.g., surveying engineering, but lacks a relation to the use of the data. In a world that is constantly changing, observations cannot be exactly repeated—an observation made later is different from the observation made before; the customary definition is usable only if these effects are ignored and thus, strictly speaking, it is only a definition for 'correctness within limits'. These limits are usually assumed such that smaller imperfections do not negatively affect the intended use. Autocorrelation (E2) is the reason that this definition of data quality works. An observation a little bit later or nearly at the same place produces nearly the same value; repeated observations would not be meaningful in a world not strongly spatially and temporally autocorrelated (Figure 4).

4.3 Quality as Result of Data Acquisition Methods

Data quality is the result of the production process of the data; this can be called "external precision". The quality of data can be deduced from properties of the production process (Timpf et al. 1996; Timpf et al. 1997). This is similar to the view that the quality of a car results from the details of the production process and thus fits general methods of industrial engineering to assess product quality (e.g., photogrammetry projects are often specified by process descriptions). The obtained characterization of data quality is derived from statistics of repeated observations obtained with the same data acquisition method.

Unfortunately, these production oriented definitions of data quality are mostly irrelevant for the use of geographic data. Practitioners resist pressure to make costly efforts to produce such data quality descriptions because they are not informative for potential users (Hunter et al. 2000).

4.4 Fitness for Use

An alternative definition is based on the concept of *fitness for use* (Chrisman 1985), which I quantified as comparison between required and provided quantity (resp. error, which is expected vs. error tolerated) (Frank 1998). Rönnbäck's critique that this gives not much guidance to the user is justified (Rönnbäck 2004) and indicates that more research is necessary (Boin et al. 2007b). Information is used to make decisions, which are then translated into actions. Decisions are the only economic use of information. A convenient definition of information reflects this: Information is an answer to a human question (Frank 1997). People ask ordinarily questions in order to make decisions, sometimes the decisions are imminent and sometimes we just collect information to be prepared for decisions we expect to take later. If information is used to make decisions, then the quality of the information can be related to the quality of the decisions made. Data that produced the information is of good quality if the resulting decision is good.

To assess the quality of the decision as the result of the quality of the data brings us back to the semantic loop (Figure 3): reality and information realm are connected (1) by observations, which populate the information realm and (2) the decisions and actions, which change the world. To assess the quality of the information one must assess the quality of the decision and how it is influenced by the information. The next section produces a highly simplified model of the decision process adapted to this end.

5 Decision Process

The ontological commitments for incomplete, uncertain, and erroneous information must be linked to the decision process to see how they affect the quality of the decisions. This requires a highly simplified summary model of how decisions are made:

The decision to take some actions starts with a goal, i.e., an imagined future world state that is desirable to the agent. For example, I am hungry and imagine a future world state in which I have eaten. I consider then a set of alternative actions to achieve that state and evaluate the different plans in order to select the best course of action, which I then carry out. Not all aspects of this model must be conscious to the agent (Roth 1994). It is sufficient that the agent selects one of the alternatives because it appears—given the current state of his knowledge—the best option. It is implied that decisions can be wrong, are made with insufficient information, etc.; the decision is sufficing and the rationality is bounded by the limitations of the agent (Simon 1956).

Note: the apparent difference between:

a) a decision whether a plan can be executed or
b) a selection of an optimal plan to achieve a goal

is not fundamentally relevant for the following discussion of influences of data quality on decision, because the first case can be logically reduced to the second one.

5.1 How and Why Decisions Can Be Wrong

A decision can be wrong in three ways:

1) The action that is decided cannot be carried out.
2) The achieved state of the world does not satisfy the goal.
3) The action did achieve the intended goal, but was not optimal.

The first two cases of wrong decisions are primarily related to the decision type (a) above—the selected action does not achieve the intended goal, whereas the third relates to the selection of one variant among several ones (case (b) above).

Effects of imperfection in the data are different if they are caused by errors in measurements (E1) or by omissions or commissions; which are implied in the same commitments by negation (P1, P2, and P3); these two types of error are discussed first for incorrect decision and then for suboptimal decisions.

5.2 Information Is Correct If It Leads to Correct Decisions

A practical definition is to state that information is correct if it leads to correct decisions. This requires first a definition of what we mean by ‘correct decision’. Let me start with a counterexample: information is incorrect if it leads to a wrong decision. For example, my decision to go to the airport at 7:30 a.m. to catch the plane for Frankfurt is in error if the plane has actually left at 7:15 a.m. Other example: my decision to buy a 2 m extension cord to connect my stereo system is incorrect if I find at home that the cable is too short because the distance between the power outlet and the stereo system is 3 m. *A decision is not correct if it does not lead to the desired goal* (i.e., flying to Frankfurt, connecting the stereo set)—this points out that decisions are taken in order to achieve a certain goal; if the action decided upon does not achieve the desired goal, the decision is incorrect. The concept of correctness assumed here is relative to the decision and not one of an unattainable perfect agreement.

If we assume bounded rationality in the decision process then the information available is influencing the decision—thus information that leads to the correct decision is correct information. Note that this definition is less influenced by the imperfections in the data, than the definition of correctness based on repeatability and takes into account the influence of error and uncertainty on the information. Much error, uncertainty, and incompleteness in the information can be tolerated as long as the action decided on achieves its goal. Statistical tests can be used on repeated observations to assess if the values obtained are within the expected margins with a certain probability. Therefore this statistical characterization of data must be connected with the decision (see (Frank 2007)).

5.2.1 Incorrect decision due to observation (measurement) error

An action is not possible as planned because of observation errors. This is the type of error surveyors have extensively studied and try to avoid with precise measurements. Most spectacular are the efforts by surveyors to assure that the two ends of a new tunnel meet in the middle of the mountain; or a surveyor measures the gap between the roads on both sides of the river and measures the steel bridge, which should fit in the gap; if the measurements are in error, closing the gap is not possible.

5.2.2 Incorrect decision due to lack of knowledge (omission or commission)

An action can be impossible because some crucial information was not available. For example driving to a city and finding that the city is on the other side of a river not shown on the map and no bridge available. A case of an instruction from a car navigation system to cross a river, where a ferry and not a bridge should be used, was widely publicized; the driver drove the car into the river and blamed the incomplete information from his navigation system (Kuhn 1994). This event demonstrates how commissions and omissions have similar effect, e.g., a commission error showing a bridge crossing a river where there is none leads to comparable errors as showing no bridge where there is one.

5.2.3 Suboptimal decision

Information present is incorrect and therefore the selected action can be executed but is not optimal for the situation; this can be caused by measurement errors, omissions, or commissions. A map shows a road as shorter than it actually is, or shows a road, which is not (yet) existing and one decides to take this shorter route, only to realize later, that this road is longer than another one or not yet ready and a longer route must be followed.

The economic effects of selecting not the optimal, but the second or third best action especially due to measurement errors, are in general not very important—because the difference between optimal choice and second or third best choice are usually not large (test this when using a car navigation system!). The small difference between alternatives is the effect of the autocorrelation already mentioned but also caused by the intentional construction of infrastructure in the world that are, whenever possible, redundant. If one fails, there is always a second option. Mankind has learned how to live in a world of error and uncertainty!

6 Influence of Imperfections of Measurement Errors

6.1 Measurement Errors

Measurement errors must be very substantial before they influence a decision. Measurement errors can

1) lead to the selection of non-optimal solution, if the values for two alternatives are closer than the error on the measurement.
2) hide that an action is not possible and lead to a decision for an unfeasible action when the measurement value and the accepted value are very close.

In actual decision situations, the first case causes a small loss in the proportional to the measurement error (Figure 9). The second case must be detected by the decision maker and judged ‘risky’; she then can change the situation to avoid the risk or reduces the error on the measurement values.

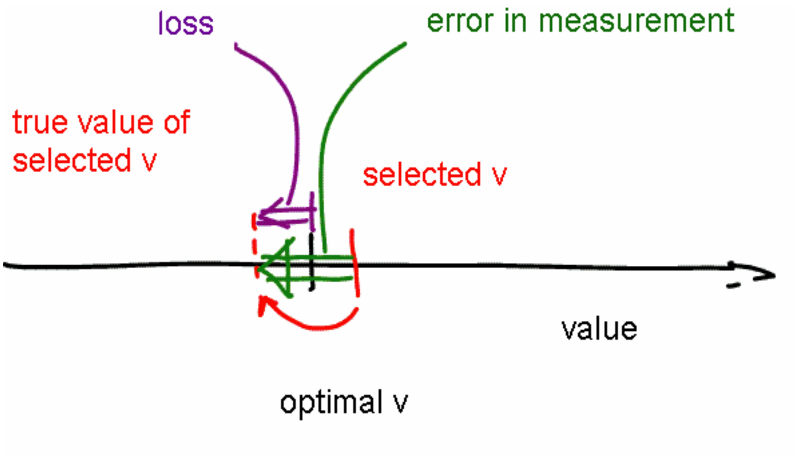

Figure 9: Loss from selecting a suboptimal solution based on measurement error

6.2 Omissions and Commissions

It is tempting to link decision errors for lack of knowledge to omissions and the selection of a not optimal action to commission, but indeed omissions and commissions produce both similar errors in decisions. For a decision error due to physical impossibility, consider a property that is crucial for the execution of the action, but this property was asserted as an effect of a commission error and is not really present—this has the same effect than the omission of a property that must be absent for the action to conclude. In such cases, omissions and commissions have the same effect, both make it impossible to execute an action that was decided upon. Selecting an action that is not optimal (suboptimal decision) can be equally the effect of an omission or a commission, and possibly even of a measurement error (Frank 2007).

7 Small Example: Errors in Road Navigation Decision

In a decision on road navigation, i.e., which road to follow to drive to another place, the three types of errors in decisions due to information quality can be demonstrated. Assume that we need to drive from *A* to *B* on a Sunday and have gas in the car for 100 km; the information we have is shown in Figure 10 left. The shortest path seems to be *x* and we decide to take it.

The data used for the decision are grossly imperfect and the true situation is shown in Figure 10 right. The decision is, firstly, to use route *x* in error due to imprecise measurement (case 1), the path *x* is very convoluted and actually 120 km long and we will fail to achieve our goal for lack of gas. The decision is, secondly, in error for lack of knowledge (case 2) because *B* is on an island and the ferry runs only on workdays (on Sundays one must take path *y*). On a workday, the decision to use route *x* would not be optimal (case 3) because the length of path *z* is not 85 km as marked but only 65 km.

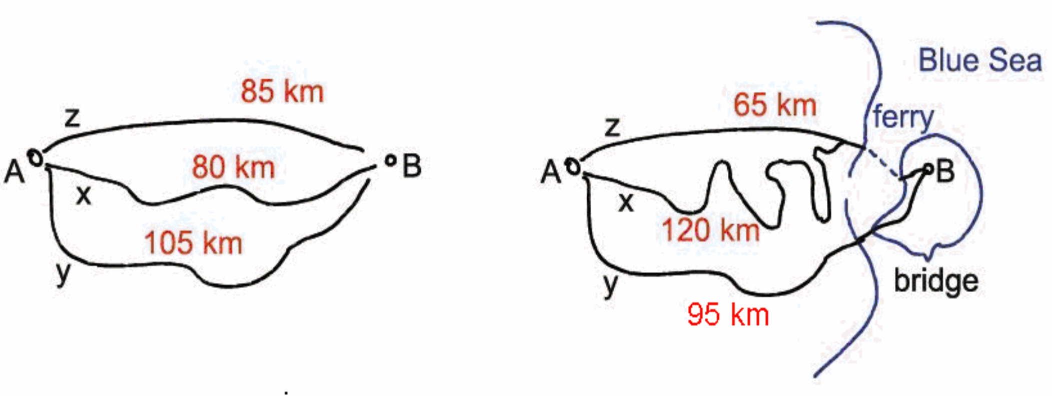

Figure 10: Information available for decision (left) and true situation (right)

The analysis of the process of identification of an optimal solution demonstrates that lack of relevant information (P 2 or P 3) and errors in object formation and measurements (R 5 to R8) connects between types of imperfections and decision quality. Errors due to classification errors are not considered here and are left for future work.

8 Conclusion

Ontologies describe conceptualization of some subset of reality (Gruber 1993); most studies of ontologies, especially research in the context of spatial information, does not consider the unavoidable imperfection in the data. This has hindered the use of ontological analysis to clarify what we understand by data quality. This article presents a novel ontology for imperfect data, or, more exactly, an ontology of imperfections in data. It was first necessary to give up the simplifying assumption that reality is linked by an isomorphism to (geographic) information system and to describe the process that connects the reality realm to the information realm—a step that was proposed earlier (Frank 2000b).

The tiered ontology (Frank 2001) leads to a systematic inventory of different types of imperfections in the data resulting from the processes of data acquisition and transformations, here presented as ontological commitments to imperfect data. The effects of imperfection in the data on decisions gives the framework in which data quality as “fitness for use” can be studied. Data quality is higher, if the data contains less imperfections, but a construction of “quality = absence of imperfections” is not practically useful. Data quality in the sense of “fitness for use” is related to a decision and means that the effects of the imperfections do not negatively affect (distort) the decision. This definition of data quality can be related to other data quality measures (e.g., internal or external accuracy).

The systematic treatment reveals opportunities for future research:

First, the approach must be extended to include effects of classification and to apply to socially constructed reality and data describing this. The analysis presented here covers only decisions in the physical realm (e.g., engineering decisions for the design of physical artifacts, for example a bridge or a building) but does not apply to decisions including legal or social aspects.

Second, the commitments to data imperfections can be related to the processes that link the ontological tiers; the commitments listed here are limitations of the processes that acquire and transform the flow of information from reality to information realm. This has not been exploited systematically here. It could answer puzzling questions that remain unanswerable in the framework constructed so far. For example, people rely extensively on relative, not absolute, information; think of your knowledge of the position of furniture in the room around you: you are aware of relative position, and ignore absolute position for ordinary decisions without negative effects. How is this formulated as an ontological commitment? Is this an imperfection in the data?

Last, but not least, ontological clarification can lead to a reduction in jargon and improved communication of data quality to the users of geographic data; empirical studies show evidence that actual users do neither use nor understand data quality information currently provided (Boin et al. 2007a).

Acknowledgement

Discussions with many students over the years have forced me to clarify my ideas; I am grateful for the many questions and arguments they have posed! The ongoing exchanges with my colleagues—especially Gerhard Navratil, Werner Kuhn, Mike Goodchild, and Barry Smith have advanced my understanding of ontology and how it applies to GIS.

References

Bittner, S. (2001). *An Agent-Based Model of Reality in a Cadastre*. Vienna, Institute for Geoinformation.

Boin, A. T. and G. J. Hunter (2007a). *Facts or Fiction: Consumer Beliefs about Spatial Data Quality*. Proceedings of the Spatial Sciences Conference (SSC 2007), Hobart, Tasmania.

Boin, A. T. and G. J. Hunter (2007b). *What Communicates Quality to the Spatial Data Consumer?* Proceedings of the 7th International Symposium on Spatial Data Quality (ISSDQ 2007), Enschede, The Netherlands.

Burrough, P. A. and A. U. Frank, Eds. (1996). *Geographic Objects with Indeterminate Boundaries*. GISDATA Series. London, Taylor & Francis.

Ceusters, W. and B. Smith (2006). *A Realism-Based Approach to the Evolution of Biomedical Ontologies*. Forthcoming in Proceedings of AMIA 2006, Washington DC.

Chrisman, N. (1985). An Interim Proposed Standard for Digital Cartographic Data Quality: Supporting Documentation. *Digital Cartographic Data Standards: An Interim*

Proposed Standard. H. Moellering. Columbus OH, National Committee for Digtial Cartographic Data Standards. **6**.

Comber, A. J., P. F. Fisher, F. Harvey, M. Gahegan and R. Wadsworth (2006). *Using Metadata to Link Uncertainty and Data Quality Assessments*. 12th International Symposium on Spatial Data Handling, Springer.

DCMI. (2006). "Dublin Core Metadata Initiative." Retrieved 25 September, 2006, from http://dublincore.org/.

Fonseca, F. T. and M. J. Egenhofer (1999). *Ontology-Driven Geographic Information Systems*. 7th ACM Symposium on Advances in Geographic Information Systems, Kansas City, MO.

Fonseca, F. T., M. J. Egenhofer, P. Agouris and G. Câmara (2002). "Using Ontologies for Integrated Geographic Information Systems." *Transactions in GIS* **6**(3): 231-57.

Fordor, J. A. (1975). *The Language of Thought*, Thomas Y. Crowell Co.

Franck, G. (2004). Mental Presence and the Temporal Present. *Brain and Being: At the Boundary between Science, Philosophy, Language and Arts*. G. G. Globus, K. H. Pribram and G. Vitiello. Amsterdam, Philadelphia, John Benjamins**:** 47-68.

Frank, A. (1983). *Datenstrukturen für Landinformationssysteme - Semantische, Topologische und Räumliche Beziehungen in Daten der Geo-Wissenschaften*. ETH Zürich.

Frank, A. (2007). *Assessing the Quality of Data with a Decision Model*. 5th International Syposium on Spatial Data Quality 2007, Enschede, NL.

Frank, A. U. (1997). Spatial Ontology: A Geographical Information Point of View. *Spatial and Temporal Reasoning*. O. Stock. Dordrecht, Kluwer**:** 135-153.

Frank, A. U. (1998). Metamodels for Data Quality Description. *Data quality in Geographic Information - From Error to Uncertainty*. R. Jeansoulin and M. Goodchild. Paris, Editions Hermès**:** 15-29.

Frank, A. U. (2000a). Communication with Maps: A Formalized Model. *Spatial Cognition II (Int. Workshop on Maps and Diagrammatical Representations of the Environment, Hamburg, August 1999)*. C. Freksa, W. Brauer, C. Habel and K. F. Wender. Berlin Heidelberg, Springer-Verlag. **1849:** 80-99.

Frank, A. U. (2000b). "Geographic Information Science: New methods and technology." *Journal of Geographical Systems, Special Issue: Spatial Analysis and GIS* **2**(1): 99-105.

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.

Frank, A. U. (2003). Ontology for Spatio-Temporal Databases. *Spatiotemporal Databases: The Chorochronos Approach*. M. Koubarakis, T. Sellis and e. al. Berlin, Springer-Verlag**:** 9-78.

Frank, A. U. (2006). *Distinctions Produce a Taxonomic Lattice: Are These the Units of Mentalese?* International Conference on Formal Ontology in Information Systems, Baltimore, Maryland, IOS Press.

Frank, A. U. (to appear 2007). *Incompleteness, Error, Approximation, and Uncertainty: An Ontological Approach to Data Quality*. Geographic Uncertainty in Environmental Security. NATO Advanced Research Workshop, Kiev, Ukraine, Springer.

Goguen, J. and D. F. Harrell. (2006). "Information Visualization and Semiotic Morphisms." Retrieved 01.09.06, 2006, from http://www.cs.ucsd.edu/users/goguen/papers/sm/vzln.html.

Gruber, T. R. (1993). *Toward Principles for the Design of Ontologies Used for Knowledge Sharing*. International Workshop on Formal Ontology, Padova, Italy., Kluwer Academic Publishers.

Henninger, W. (2007). *ATKIS® in Bayern und seine Folgeprodukte*. 14. International Geodätische Woche Obergurgl 2007, Obergurgl, Wichmann.

Hunter, G. J. and E. G. Masters (2000). *What's Wrong with Data Quality Information?* Proceedings of the GIScience 2000 Conference, Savannah, Georgia, University of California Regents.

Jackendoff, R., P. Bloom and K. Wynn, Eds. (2002). *Language, Logic, and Concepts*, MIT Press.

Karmiloff-Smith, A. (1994). "Reprint: Precis of: Beyond Modularity A Developmental Perspective on Cognitive Science." *Behavioral and Brain Sciences* **17**(4): 27 (693-745).

Karmiloff-Smith, A. (1995). *Beyond Modularity A Developmental Perspective on Cognitive Science*. Cambridge, MIT Press.

Kent, W. (1979). *Data and Reality Basic Assumptions in Data Processing Reconsidered*. Amsterdam, New York, Oxford, North-Holland Publishing Company.

Kuhn, W. (1993). Metaphors Create Theories for Users. *Spatial Information Theory*. A. U. Frank and I. Campari, Springer. **716:** 366-376.

Kuhn, W. (1994). Zur Verwendbarkeit von Geographischen Informationssystemen (GIS). *UTA (Umwelt Technologie Aktuell)*. **5:** 88-93.

Kuhn, W. (Draft 2007). An Image-Schematic Account of Spatial Categories. Münster, Germany, Institute for Geoinformatics, University of Münster**:** 12.

Kuhn, W. and A. U. Frank (1991). A Formalization of Metaphors and Image-Schemas in User Interfaces. *Cognitive and Linguistic Aspects of Geographic Space*. D. M. Mark and A. U. Frank. Dordrecht, The Netherlands, Kluwer Academic Publishers**:** 419-434.

Lawvere, F. W. (1965). Algebraic Theories, Algebraic Categories, and Algebraic Functors. *Theory of Models*. Amsterdam, North-Holland**:** 413-418.

Mac Lane, S. (1998). *Categories for the Working Mathematician*. New York, Berlin, Springer.

Masolo, C., S. Borgo, A. Gangemi, N. Guarino and A. Oltramari (2003). WonderWeb Deliverable D18 (Ontology Library). Trento, Italy, Laboratory For Applied Ontology - ISTC-CNR**:** 247.

McCarthy, J. and P. J. Hayes (1969). Some Philosophical Problems from the Standpoint of Artificial Intelligence. *Machine Intelligence 4*. B. Meltzer and D. Michie. Edinburgh, Edinburgh University Press**:** 463-502.

Navratil, G. and A. Frank (2006). *What Does Data Quality Mean? An Ontological Framework*. AGIT 2006, Salzburg, Wichmann Verlag.

OGC. (2000). "The Open GIS Consortium Web Page." Retrieved 21 November, 2000, from http://www.opengis.org.

Rönnbäck, B.-I. (2004). *Are Uncertain Uncertainties Useful? Towards Improved Quality Assessment of Spatial Data*. Luleå University of Technology. PhD.

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.

Roth, G. (1994). *Das Gehirn und seine Wirklichkeit*. Frankfurt am Main, Suhrkamp Verlag.

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

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

Smith, B. and P. Grenon (2004). "SNAP and SPAN: Towards Dynamic Spatial Ontology." *Spatial Cognition and Computing*(4): 69-103.

Tarski, A. (1944). "The Semantic Conception of Truth and the Foundations of Semantics." *Philosophy and Phenomenological Research* **4**.

Timpf, S. and A. U. Frank (1997). "Metadaten - vom Datenfriedhof zur multimedialen Datenbank." *Nachrichten aus dem Karten- und Vermessungswesen* **Reihe I**(117): 115-123.

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.

Wand, Y. and R. Y. Wang (1996). "Anchoring Data Quality Dimensions in Ontological Foundations." *Communications of the ACM* **39**(11): 86-95.