INCOMPLETENESS, ERROR, APPROXIMATION, AND UNCERTAINTY: AN ONTOLOGICAL APPROACH TO DATA QUALITY

Ontology for Data Quality

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

Abstract. Ontology for geographic information is assumed to contribute to the design of GIS and to improve usability. Most contributions consider an ideal world where information is complete and without error. This article investigates the effects of incompleteness, error, approximation, and uncertainty in geographic information on the design of a GIS restricted to description of physical reality. The discussion is organized around ontological commitments, first listing the standard assumptions for a realist approach to the design of an information system and then investigating the effects of the limitations in observation methods and the necessary incompleteness of information. The major contribution of the article is to replace the not testable definition of data quality as 'corresponding to reality' by an operational definition of data quality with respect to a decision. I argue that error, uncertainty, and incompleteness are necessary and important aspects of how humans organized and use their knowledge; it is recommended to take them into account when designing and using GIS.

Keywords: Incompleteness, Error, Approximation, Uncertainty, Error Ontology, Spatial Ontology, Spatial Data Quality

*Andrew U. Frank, Institute for Geoinformation and Cartography, Technical University Vienna, Gusshausstrasse 27-29/127, A-1040 Vienna, Austria; email: frank@geoinfo.tuwien.ac.at

1. Background

The goal of human activities is to improve one's situation and—following the Golden Rule—to improve the 'condition humaine' in general. This is part of a Greco-Judaic tradition to control the world and use it (Genesis 1, 28). Information became central for the development of economy in the past few centuries. The industrial revolution in the 18th and 19th century improved on the production of goods for human consumption and allowed an unprecedented increase in population; it combined improvements in government, taxation, and markets together with technical improvements in manufacturing (North 1981). North identifies a second economic revolution when scientific methods are used to produce systematically new knowledge to further advance technology and management. This is evident in the current debate on directing universities to produce 'socially useful and responsible knowledge' combined with high levels of funding for universities but it is equally true for all the new internet businesses. Information has become a factor of production, comparable to the classical production factors of land, capital, and labor (Ricardo 1817; reprint 1996; Marx 1867; translated reprint 1992; Frank to appear 2005).

If information is a production factor like others, it must be measurable both in quantity and quality. Efforts to include "knowledge" in the accounting of large companies are under way (Schneider 1996), but problems of measuring quantity and quality remain. Easily observable and countable substitutes (number of patents, number of scientific publications, etc.), which are expected to be proportional to the actual knowledge, are often used. I have suggested a method to measure the quantity of pragmatic (useful) information (Frank 2001; Frank 2003b), but the approach is currently viable on a micro level only.

What do we mean when we say that information is of high quality? Before the computer age, one would have said 'the information is from reliable sources', qualifying the information not directly but indirectly by its source. In today's information economy, quality of information becomes important for business. The loss for U.S. businesses due to data quality problems is estimated as $600 billion for 2002 (Eckerson 2006).

Quality of information is a novel concept, which has not been used before; scientists—especially astronomers and surveyors—commented on the quality of observations in the 18th century; surveyors have generalized this approach to evaluate the precision of observations and contributed to the data quality discussion (Chrisman 1985; Frank 1990) (Robinson et al. 1985). Business processes using data go astray if the inputs are wrong and this gives an alternative approach to the topic (Wand et al. 1996).

Quality of information is even more important today in the transformation of the economy from maximum production at any cost to an economy respecting 'limits of growth' (Meadows et al. 1972; Pestel 1989). Mitigation of environmental disasters like flooding, tsunami, or forest fires are political and economical goals making detailed and high quality information necessary for their achievements. Mankind has learned that not everything that is technically possible is desirable. We need to understand the laws of environment as well as we understand the laws of physics: the construction of a perpetuum mobile (perpetual motion machine) is attempted today only by fools, because we understand the inviolable laws of thermodynamics. Plans affecting the environment too often violate environmental laws that are equally forceful; we find increasingly that 'the environment kicks back' when we ignore its rules. Information, especially spatial information, plays the crucial role to understand and eventually use, to our advantage, the laws of environment.

2. Goal

In this article I want to link the methods used to collect, manage, and use environmental data with the ontological commitment, which are tacitly assumed. This seems useful to avoid that contradictions in the assumptions lead to confusion and inappropriate use of, in principle valid, methods. Identifying the ontological commitment is important in today's complex and diversified edifice of science to achieve consistency across different disciplines and applications. The focus is on data quality; the connection between data quality and ontology has been made before (Wand et al. 1996) and I hope to extend this original contribution in a way different from recent papers by (Ceusters et al. to appear 2006). The paper is restricted to descriptions of the physical reality and the extension of the arguments to cultural aspects, e.g., political boundaries, land ownership, etc. is left for future work.

The goal of Ontology in philosophy is to understand how the world is and how things exist in the world. It starts with Aristotle's Metaphysics (Aristotle 1999). The term ontology was created in the 19th century; foundational contributions were made by Husserl. The difficulty with the philosophical tradition of Ontology is that human knowledge is limited to what we can observe; phenomenology concentrated on the limits of our abilities to know about the world (Bergson 1896; reprint 1999). The movement of existentialism (Heidegger 1927; reprint 1993; Sartre 1943; translated reprint 1993) contrasts with analytical philosophy using increasingly formal methods, coinciding with foundational questions in mathematics (Whitehead et al. 1910-1913).

Ontology, as produced by philosophers, tries to give a consistent description of the world and how it exists in general. It is useful to identify the assumptions and point out inconsistencies but logical deduction alone cannot tell us 'how the world is'. It can, at best, demonstrate that some set of assumptions are not consistent. Knowledge about the world is only achieved starting with observations and is thus dependent on the world *and* on the observation system.

Practical use of ontology—with a lower case o—in information systems has goals that are more modest: it gives rules how consistent descriptions of conceptualizations of a subset of reality for a purpose (Gruber 2005). Any information system has an underlying ontology—the conceptualization of the part of the world, which is included—even if it is not described explicitly. Designers and users of an information system construct a mental model of the subset of the world they are interested in and communicate this model verbally; such symbolic representations are then entered in an information system. The data structure of the information system is a representation of this conceptualization; if it is described in a formal ontology (Smith 1998) then the description can be analyzed and inconsistencies detected and resolved.

In this article, I want to explore the ontological commitments, which are necessary for an information system in a realistic view, i.e., a view that takes into account error, approximation, and uncertainty. The goal is to give a consistent set of ontological commitments, which allow a definition of data quality and how it is practically used. The focus is on geographic information systems and how they are used in environmental applications—but the results should be fully general for information systems independent of purpose. This approach is different from Wand and Wang's effort: they considered primarily questions of mapping between facts and their symbolic representation and assumed that the granularity of the representation is properly set; here I want to explore the difficulties that result from granularity and differences in the granularity when merging multiple data sets.

3. Ontological Commitments

Avoiding ontological commitments is not possible—designers of information systems can only avoid making explicit how they conceptualize the subset of the world they are modeling in the system. Making their choices explicit avoids inconsistencies, improves communication among multiple designers and eventually communicates the conceptualization to the users of the system, again avoiding misunderstanding and misuse of the

information provided by the system (Fonseca et al. 1999). In this paper, I expand this view, which is common today, to include data quality aspects.

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. Few philosophers and many writers of science fiction (Asimov 1957; Adams 2002) have explored the logical consequences of constructions in which either no world outside of my thinking exists (Schopenhauer 1819 & 1844; translated reprint 1966) or in which multiple worlds coexist. They lead to inconsistencies, which often provide for interesting reading, but not to an account of the world as we experience it.

3.2. COMMITMENT O 2: THE WORLD HAS EVOLVING STATES

The world has states, which evolve in time. This ontological commitment is twofold: it posits a single time and changeable states of the world (this is postulate 1 of Wand and Wang (1996, 89)).

3.3. COMMITMENT O 3: OBSERVABLE AND CHANGEABLE 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)). Observation of physical state for certain properties and a point is objectively possible (point observations); the influence of the observer on the observation value is small and repeated observations give the same values.

An extensive discussion of the influence of the observer on the observation has been carried out in the social sciences, where subjective judgments of situations are heavily influenced by the background of the observer and in physics, where observation influences the state of the observable (Leinfellner 1978; Mittelstraß 2003). These difficulties do not affect the treatment here: the observable states of the world are restricted to states of the physical (macro-) reality as they are measured with standard measurement devices and the result expressed in SI units or similar (e.g., cm, g, s). I exclude from this discussion (a) physiological states of individuals, as discussed in measurement sciences (Krantz et al. 1971), (b) assessment of cultural reality are included in the observable states, nor (c) quantum physics. The observable states of importance in a GIS are within the realm of classical physics and do not include quantum effects.

The actors in the world can not only observe the world, but they can also affect changes in reality 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 2) that connects the observations with their sensors to the changes with their actuators and combines the semantics of observations with the semantics of changing operations (Frank 2003a).

3.4. COMMITMENT O 4: INFORMATION SYSTEMS ARE MODELS OF REALITY

Observation results in information and we have to discuss both the system of reality and the information system (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 1). The information realm is a partial and incomplete model of the world, somewhat as described by Plato in his cave metaphor. By model, we understand a structure, which is related by a morphism with the world. Corresponding operations in the model have corresponding results (Kuhn et al. 1991; Goguen et al. 2006). The focus of Wang and Wand is this mapping, which they characterize along the same lines as customary in category theory (Asperti et al. 1991), (similar recently (Ceusters et al. 2006)).

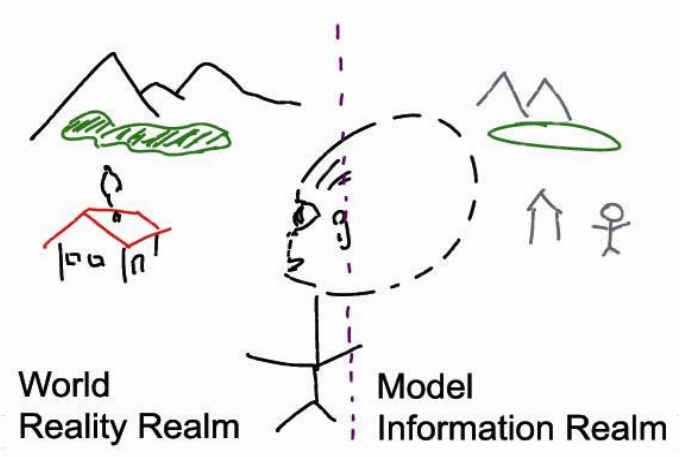

Figure 1: The Reality Realm and the Information Realm

The division of the ontology in a world realm and an information realm is an important contribution of Wand and Wang; the two realms are related by a morphism. The quality of the information is threatened if the relation between the things in the world and the entities in the information realm are not isomorphic (i.e., one-one). If the two realms are linked by an isomorphism, which is often assumed, then the distinction between thing in

the world and in the information model would not be required from an algebraic point of view. Reality and the model are the same—up to isomorphism (Mac Lane et al. 1991). Actual information systems do not achieve an isomorphism: sometimes one thing in the world corresponds to multiple entities in the information model, sometimes situations in the world have no related representation in the information model; in most cases, 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 be glossed over (Kent 1979).

3.5. COMMITMENT O 5: SEPARATE PHYSICAL AND INFORMATION CAUSATION

The changes in the state of the world are modeled by physical laws: The cause for water flowing downward is gravity, the cause of a bullet to fly are the forces resulting from a chemical reaction, when the explosive in the shell is ignited. The rules of physics can be modeled in the information realm and allows the construction of future states in the information realm. This is extensively used to predict what the effects of actions are and the foundation of all planning. The change in the physical world can be modeled as a Markov chain—a following state is the result of the current state or of the current state and previous states.

A second and entirely different form of causation, which I will call *information causation,* is the result of decisions by agents. Agents have *free will* and can make decisions about their actions (Searle 2001). Decisions are in the information realm but they affect—through physical laws—the reality realm. In a macroscopic view, a successor state is independent of the previous state of the world.

It is, however, important to note that decisions can have the intended effect only if—and only if—the action can be carried out and no physical laws contradict it. For example, deciding to move from Vienna to Kiev in one hour by car is possible, but the decision cannot be carried out because several physical (and traffic) laws prohibit my car to drive at the necessary speed of 1000 km/h, etc. Despite my decision, the desired result cannot be achieved, because I cannot start a chain of physical causations to realize my decision; one could say that the mapping of the information causation from the information realm to the reality realm does not exist in this case.

4. Quality of Information

How to define quality of information in this context? How to give more content to the idea that given information is of high quality if it corresponds with reality? Indeed, what is meant by ‘correspond’?

The most often used definition of information quality is based on the repeatability of observations. Assuming that a state of the world has not changed, then an information is correct in this sense that if the information is the same as obtained by a new observation. This is the definition used by Wang and Wand. In their contribution, they point out that not only the recorded information must be considered, but also other information that can be inferred (definition 2 and postulate 6). This definition assumes that the quality of the information is—at least in some dimensions—fixed appropriately with respect to the intended use. 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, only a definition for ‘correctness with some limits’.

This more traditional approaches to data quality is often used, because it considers the production of the data and deduces the quality of data as properties resulting from the production process (Timpf et al. 1996; Timpf et al. 1997; Timpf 2002). 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 to assess product quality. Unfortunately, these definitions of data quality are mostly irrelevant for geographic data and its use. Practitioners resist to use data quality descriptions following current standards (Hunter et al. 2000) because they are not informative for potential users.

An alternative definition is based on the concept of ‘fitness for use’ (Chrisman 1985). The information is used to make decisions, which are then translated into actions. This is the only use of information and is reflected in a convenient definition of information: Information is an answer to a human question (Frank 1997). People ask questions in order to make decisions, sometimes the decisions are imminent and sometimes we just collect information to be prepared for later decisions. If information is used to make decisions, then the quality of the information can be related to the quality of the decisions made.

To assess the quality of the decision brings us back to the semantic loop: reality and information realm are connected (1) by observations, which populate the information realm and (2) the decisions and actions, which change the world (Figure 2). To assess the quality of the information one must assess the quality of the decision and how it is influenced by the

information. The connection between data semantics and quality is revealed with this viewpoint.

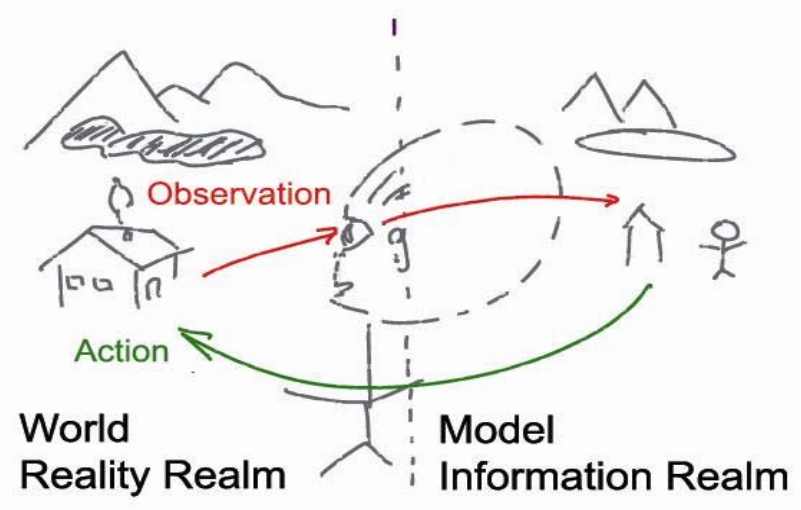

Figure 2: Closed Loop Semantics connect Reality Realm with Information Realm through observations and actions

5. Ontology of Error and Uncertainty

The ontological commitments were listed above primarily to remind the readers of what is implied in an information system design; they are the usual assumptions underlying the construction of geographic information systems (Frank to appear) and are the justification for the ontology-driven approach to design an information system (Fonseca et al. 1999). These "usual' ontological commitments are unrealistic, because they ignore error and uncertainty in our knowledge of the world, and pretend that we have perfect knowledge. This illusion is acceptable given that we have most of the time sufficient information to function at acceptable performance levels in our environment but it is not usable to construct more advanced information systems, which use data collected for other purposes and following different quality standards; in such combinations of data from different sources, we must take into account the limitations in our knowledge. Understanding error and uncertainty in data is therefore crucial to achieve interoperability of geographic data collections (Vckovski 1997).

5.1. COMMITMENT EU 1: INFORMATION ABOUT THE WORLD IS INCOMPLETE

The world is infinitely complex and the information we have about it is always limited. 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 we construct is therefore always limited in the level of detail and the completeness of the aspects modeled; most of what is in the world must be left out; our models are restricted to the aspects that are relevant for the decisions we intend to make. The level of detail is linked to the purpose of the GIS and the decision expected to be made with the information. This contradicts the optimistic view of GIS as a single 'multi-purpose' or even 'all-purpose' spatial information system in the early days (Gurda et al. 1987).

5.2. COMMITMENT EU 2: OBSERVATIONS ARE ERRONEOUS

Observations of the changeable states of the world are never perfect. They are affected by unavoidable effects, which create differences between the ideal observation (the ideal *true value*) and the actual realization of the observation. These effects can be random and are often modeled by normal distributions. Observations are also affected by systematic effects, e.g., a yardstick is too short or a watch runs always slow. Such systematic effects can be controlled and eliminated by observation methods, but random disturbances cannot be avoided and affect all observations of physical properties.

5.3. COMMITMENT EU 3: AUTOCORRELATION

One might ask how people survive in a world where the information we have is necessarily incomplete and erroneous. To conclude that goal directed actions and survival is impossible, would be premature ("Philosophers should be very careful when they deny the obvious" Searle). But what counteracts the effects of the fact that all our knowledge is incomplete and erroneous?

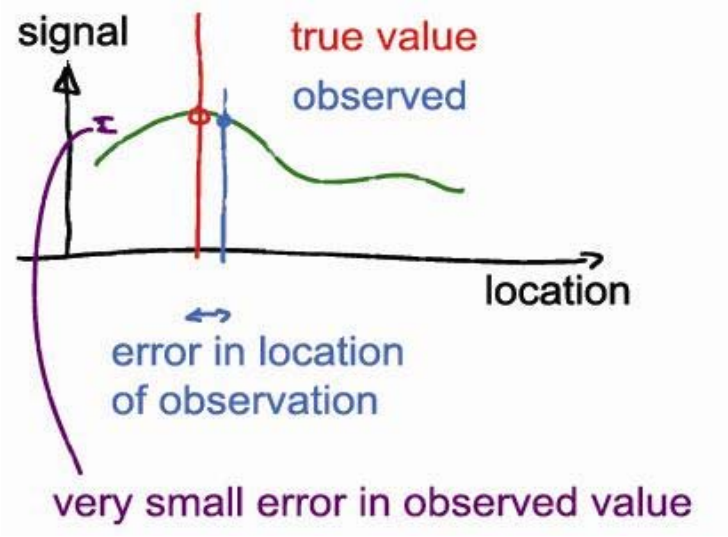

Figure 3: Large error in location leads to small error in observed value

The physical world is strongly autocorrelated—both in space and time. 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. Correlation between different observations is also strongly correlated: for example sugar content of a fruit and its color is often correlated—we pick the red strawberries, because they are sweet and taste good and leave the green ones. The strong autocorrelation is also at the base that the usual definition of data quality as 'corresponding to reality' 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 3).

Life in a world without the strong spatial and temporal autocorrelation would be very difficult if not impossible. Most of the world is slowly and continuously changing and we focus on the discontinuities. On the background of stability we focus on the interesting changing and discontinuous points.

5.4. COMMITMENT EU 4: BIOLOGICAL AGENTS HAVE LIMITED INFORMATION PROCESSING ABILITIES

The structure of our information is not only influenced by reality but also by the systems to process information. The abilities of the brains of biological agents—including humans—are very limited and the biological, i.e., energy, cost of information processing, is high. Biological agents have therefore developed methods to reduce the load on their information processing systems—commonly called the 'cognitive load'—to allow efficient decision making with limited effort and often in short time.

5.5. COMMITMENT EU 5: OBJECT CENTERED DATA PROCESSING

Processing of information describing reality is primarily object oriented. Humans, and many other biological agents, structure the observations they perceive in information about objects. The observation of properties of points in space and time are restructured to become properties of objects. Objects are constructed such that they endure in time and have constant properties over time. Spatial and temporal autocorrelation makes this reduction of cognitive load possible.

The cognitive system forms objects at boundaries of continuities and reduces therewith the cognitive load: it is simpler to keep track of objects with uniform and seldom changing 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. Autocorrelation is similarly used in technical systems to reduce bandwidth necessary for transmission, e.g., of television images; it is the reason data compression methods like JPEG and MPEG work.

Our cognitive system is so effective because it identifies objects in the array of sensed values, and we reason with objects and their properties, not with the multitude of values sensed. Thinking of tables and books and people is much more effective than seeing the world as consisting of data values for sets of cells. It is economical to store properties of objects and not deal with individual raster cells. We cut the world in objects that are meaningful for our interactions with the world. As John McCarthy has pointed out:

> "…suppose a pair of Martians observe the situation in a room. One Martian analyzes it as a collection of interacting people as we do, but the second Martian groups all the heads together into one subautomaton and all the bodies into another. .. How is the first Martian to convince the second that his representation is to be preferred? He would argue that the interaction between the head and the body of the same person is

> closer than the interaction between the different heads ... when the meeting is over, the heads will stop interacting with each other but will continue to interact with their respective bodies." (McCarthy et al. 1969, 33).

Our experience in interacting with the world has taught us appropriate subdivisions of continuous reality into individual objects. Instead of reasoning with arrays of connected cells, as it is done in finite element analysis for, e.g., strain analysis or movements of oil spill, reasoning is performed with individual objects: The elements on the tabletop (Figure 4) are divided in objects at the boundaries where cohesion between cells is low; a spoon consists of all the material that moves with the object when I pick it up and move it to a different location.

Figure 4: Typical objects from tabletop space

Humans conceptualize themselves and the rest of the reality preferably in terms of objects and their properties. Objects endure in time, they have an identity and changeable state. The changeable state of objects is the consequence of the assumption that the world has changeable states and objects are aggregates of real world points. Object properties describe the state of objects; they are typically integrals over the volume the object forms (Eq1). This is usually tacitly assumed (e.g., in Wand and Wang) but creates ontological difficulties:

- object formation is not unique: different persons and for different purposes the same part of reality can be split in different ways into objects.
- The formation of object introduces error and uncertainty in the data

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

5.6. COMMITMENT EU 6: MULTIPLE WAYS TO FORM OBJECTS

Aristotle discussed familiar objects in terms of natural kinds—the classes of objects that are naturally distinct: cats, dogs, etc. There is little doubt how to form such objects and how to classify them for the natural species, because there exist hardly any borderline cases—there are no breeds between dogs and cats (but not all cases are as simple: horses and donkeys breed and produce mules). The task of the philosopher is to cut up nature at its joints. An object is considered to move as a single unit: a glass, a plate, a cat. All that moves with the object is part of the object—and only exceptionally one asks question like ‘are loose hair in the fur of an animal part of the animal or not?’ (Figure 5).

Figure 5: Three girls combing a big dog, making the boundaries of the dog sharper

Object formation is however not as simple when we consider geographic space: there are multiple ways to subdivide space into objects. Considering the terrain, we can focus on form, and identify watersheds, valleys, and mountains, but focusing on land cover, we identify fields and forests. Many other ways to subdivide space are used: ethical and religious boundaries are often debated and sometimes lead to wars. For a geographic information system, we must accept that not a single “natural” subdivision of space exists but different purposes require different approaches; a GIS must be prepared to have coexistent, overlapping spatial objects.

5.7. COMMITMENT EU 7: OBJECTS AS REGIONS WITH UNIFORM PROPERTIES

A very general approach to define objects is to say that they form regions with at least one property having a uniform value. The prototypical object—an animal—is then a connected area of cells with the same DNA; other objects are uniform in material, color, movement, etc. Solid objects, where boundaries are revealed when we move them, have uniformity in material properties that makes them 'hang together'.

For properties with a continuous value: The uniformity of the property means to be within some limits and introduces thresholds for which the property is considered uniform. This absorbs uncertainty in the observations but introduces uncertainty in the boundary of the object.

Different objects result if we select different uniform properties. Areas of uniform land cover (e.g., grass land, forest) do not necessarily coincide with watersheds and produce different objects. The autocorrelation in space and the correlation between factors influencing natural processes result in object boundaries that often (nearly) coincide. It is not by accident that the land cover on one side of the fence is different than on the other side and that the boundary of 'my garden' and the neighbor's field coincide with the fence.

5.8. COMMITMENT EU 8: OBJECT FORMATION IS UNCERTAIN

Objects are delimited by boundaries and these boundaries have observational error; a general model of objects defines them as areas of uniform values in some property. The error in observing the property value affects the determination of the boundary (Figure 6).

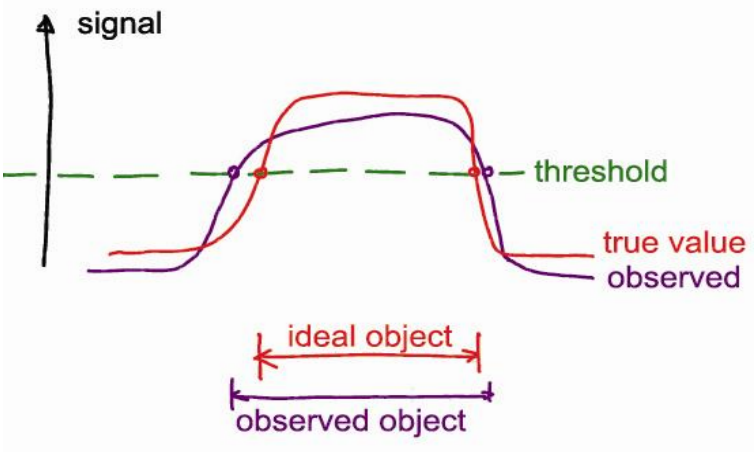

Figure 6: Error in observation of property results in error of object boundary

The objects have states that derive from the observations of point properties (commitment O 3). The properties of objects are sums (integrals) over some functions of point observations. The error in the observation of properties of the objects is therefore affected by observational errors in multiple forms:

- error in the area,
- error in the observation.

These errors can be modeled if the observation errors are assumed to be random, normally distributed (Navratil et al. 2006). However, such simplifying assumptions that are necessary to achieve a tractable formalization are unrealistic as they leave out the influence of correlations. I suggest using the term *approximation* for the difference between the true, intended value and the value resulting from observations of properties of objects.

5.9. COMMITMENT EU 9: UNCERTAINTY IN CLASSIFICATION

Objects are not just formed and described, but the formation and the description is detail to the classification of a phenomena as an object of a certain type. Objects are first instances of a class—even if this is only the most general class *Thing*—and then boundaries and properties are observed. This classification of an object has some problems that affect the quality of the data as we will see after a brief discussion of the concept of *class* (also known as *universal*, *type*, etc.) and how it is used in decision making.

The classification asserts that the object in case is part of a group of objects—the class—that share some properties. There are two ways classes are defined. In the extensional understanding of a class it is a set of objects with common properties; the intensional definition of class starts with the properties of an object (or its intended use) and the class is all objects (existing, having existed, or existing in the future) with these properties. One can imagine an ideal member of this class—the prototypical dog, mountain, etc., which is imagined as ideal universal, akin to the ideal circle; philosophers debate how such universals exist. The practical problem for information systems is that different definitions of classes are used, but described by the same word.

The descriptive terms for classes (forest, lake, etc.) are often polysemous—there are multiple concepts described by the same term. In Austria, the word ‘forest’ is used with different meanings, some of them apply even when no trees are present (but also, some terrains with trees do not classify for ‘forest’ in a legal sense).

Classification is further complicated by the so-called ‘prototype effects’ in natural language classification: not all objects in a class have some properties in common (Rosch 1978). Take the example of the class ‘bird’; one would commonly assume that birds are animals that fly—just to be reminded that also ostriches, emus, and penguins are birds, which cannot fly. Some exemplars are just better ‘birds’ than others. This applies equally to land use and land cover classifications; the prototypical forest in central Europe (the ‘dunkle Tann’ occurring in Grimm’s fairy tales) is different from what a Greek or a Finn calls a forest.

Classifications evolve in time with advances in science (Fleck 1935; reprint 1980) or with changes in the social interest. For example, land cover definitions evolved in time and the observations made based on a previous classification are incommensurable to observations with the new classification (Comber et al. 2004).

Classification is very important for human communication: we speak of cats and dogs and mean the classes of animals that have particular properties, e.g., size, form, behavior. Classifying an object based on its visible properties leads us often to assume that the object has the values typically for objects of this class for properties that we cannot observe; for example, if we classify an animal as a dog based on its visual appearance we will assume that it barks (and be very surprised if it starts to meow). Classification is thus the base of ‘default reasoning’ when we do not have particular information about the individual we assume that the usual properties of the class apply.

The uncertainty in classification comes from multiple sources, including at least:

- Selection of the property, which is uniform in the object;
- Selection of the thresholds for uniformity;
- Error in the position of the boundary;
- Errors in the observations relevant for the subclassification of an object.

Here an example: for land use classification, the property that must be uniform for an object is the land use (not land cover—but given that land cover is easier to observe, most classifications of land use are actually classification of land cover). Depending on the scale of our mapping efforts, wider or narrower thresholds for ‘uniformity’ are set: how much weeds may grow in a corn field before we stop classifying it uniformly as ‘corn’. How fine are the subdivisions for land use: agricultural (versus forest), field (versus pasture), corn field (versus wheat field). Once we have settled on corn fields and set the thresholds for weeds, the boundary of the field must be determined and measured. If we then further classify in corn fields of high yield and corn fields of low yield, an estimate of the yield is necessary.

Under the assumption that a classification process groups objects based on some determined properties in groups the uncertainty in the classification would be only from the error in the observation of properties. The formation of objects involves the uncertainty of the boundary and the errors in property observation. The approximation in the object property translates to an uncertainty in the classification. If a more reliable and precise classification is available, then the quality of a given classification can be assessed and the percentage of omissions and commissions established or a matrix of misclassification of multiple classes given. The difficulty is however more often in the imprecise or changed definition of the classes, which makes object formation and classification nearly impossible to compare (Comber et al. 2004).

The uncertainties in classification are multiple and poorly understood. Many ontologists posit that classes with fixed definitions exist, ignoring that many of the usability problems of information systems originate in differences in the classifications used during data collection and data use. I have suggested that properties of objects are used as primitive notions (and not classes as usual in taxonomies) and that classifications are defined in terms of object properties; this results in very fine grained classifications and defined rules of inference between classes (Frank to appear 2006b; Frank to appear 2006a); the idea is related to Formal Concept Analysis (Burmeister 2003).

6. Decision Process

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

The decision to take some actions starts with a goal, 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—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 and that decisions are made with insufficient information, etc. The decision is sufficing and the rationality is bounded by the limitations of the agent (Simon 1956).

7. Correct Decisions Derive from the Quality of the Information

Information cannot be correct in the sense of correspondence with reality (commitment EU 1 and 2): a repeated observation is never giving exactly the same result; the random effects and the changes in reality produce different values every time the observation is repeated. Statistical tests can be used to assess if the new value obtained is within the expected margins with a certain probability.

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 am. Other example: my decision to buy 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 plug 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 lead to the desired goal, the decision is incorrect.

If we assume (bounded) rationality in the decision process, the information available is influencing the decision—thus information that leads to the correct decision is correct information. Note that this definition does require much less, 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. A decision can be wrong in multiple ways:

- The action that is decided cannot be carried out.
- The achieved state of the world does not satisfy the goal.
- The action was not optimal; if the information would be better, another action would have been selected.

It appears useful to analyze these different reasons for actions to fail the achieved goal:

7.1. PHYSICAL IMPOSSIBILITY DUE TO OBSERVATION (MEASUREMENT) ERROR

An action is not possible because of observation errors. This is the type of error extensively studied by surveyors: Most spectacular are the measurements taken to assure that the two ends of a tunnel meet in the

middle of the mountain. Similar cases of careful measurement, 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 bridge is too long or too short, closing the gap is not possible.

In general, humans have found methods to avoid such costly and difficult measurements that have always some error. Carpenters traditionally put the beams together, cut and bore the holes at once through multiple layers and thus assure that the pieces will fit when installed in the roof—all without measurement! If a cable of a certain length is necessary most people do not try to cut to measure but make it longer—it will fit certainly, even if measurement errors are relatively large (I should have bought a 5m extension cord—it would have achieved my goal with a small additional cost!).

Many such techniques have been devised over the millennia of carpentry, tailoring, etc. to reduce the need for exact measurement; most trades avoid measurements completely! Only few situations make surveying and exact measurement necessary, e.g., the reconstruction of a boundary after it is lost due to flooding in Ancient Egypt. Measurements are necessary, when there is no 'sure side' where error does not matter: A cable can be too long without problem, a box can be too large to pack an object, but some problems have no 'secure side'—too long or too short is equally bad. For example, cooking pasta or baking bread requires exact timing—but again the goal is achieved by repeated testing and not by accurate measurement. Advanced technology increases the need for accurate measurement and planning—sea navigation, building construction with accurate planning of the forces in the building and reduced, slender pillars and many similar modern examples are only possible with accurate measurement and precision of measurement are taken into account in the design.

7.2. PHYSICAL IMPOSSIBILITY DUE TO LACK OF KNOWLEDGE

An action can be impossible because some crucial information was not available. For example driving to a city and finding out that the city is on the other side of a river or on an island—in both cases a means to cross the river (a bridge, a ferry) is required. A case of an instruction from a car navigation system to cross a river, where a ferry should be used and was not present was widely publicized, because the driver drove the car into the river and blamed the incomplete information from his navigation system (Raubal et al. 2004). This case of omission is of great importance and it is much more difficult to guard against it; Grice with his conversational implication studied information and decisions in the context of a exchange

between people, but the theory is applicable to the information we gain from consulting a database (Grice 1989).

7.3. THE ACTION SELECTED IS NOT OPTIMAL

Information present is incorrect and therefore the selected action is not optimal for the situation; this is often a case of a commission error: a map shows a road, which is not (yet) existing and one decides on a short route, which later is discovered to be longer than another route.

The economic effects are in general not very important—because the difference between optimal choice and second, third best choice are not large. This is an effect of the autocorrelation already mentioned but also part of 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!

7.4. 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 explained. Assume that we need to drive on a Sunday from *A* to *B* and have gas in the car for 100 km; the information we have is shown in Figure 7 (left). The shortest path seems to be *x*. This decision is in error due to imprecise measurement, if the path *x* is very convoluted and actually 120 km long and we will fail to reach our goal. The decision to follow path *z* is in error for lack of knowledge that *B* is on an island and the ferry runs only at workdays (on Sundays one should take path *y*). The decision to take path *x* is not optimal if we find out that the length of path *z* is not 85 km as marked but only 65 km; it would have been a better decision to take *z* and not *x*.

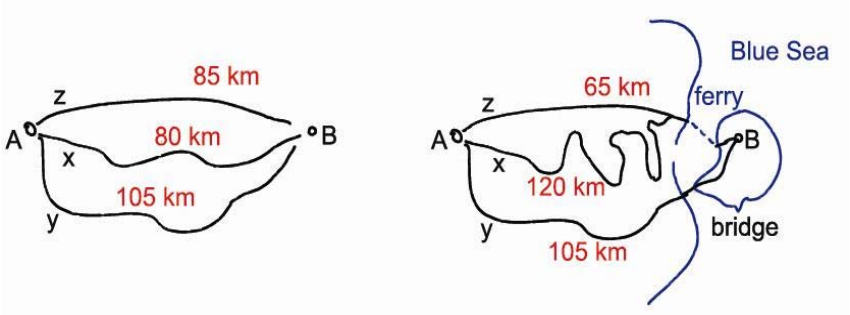

Figure 7: Information available for decision and true situation

8. Conclusion

The economic effects of measurement errors and commissions are often not very important, but errors of omission are difficult to counteract and have substantial cost. This may give a partial reason why people collect information ‘just in case’. Who has not a large library and reads all papers published in the hope that the data obtained may be useful one day? In general, the information we have is sufficient for the decisions we must make and information errors are not very costly, but often we lack the information necessary completely.

Geographic data used for administrative decision making is usually collected with proper levels of quality to make the intended decisions “reasonably well”. By reasonably well I mean that an optimum is reached between the cost of improved data quality through more efforts when collecting the data and the cost of correcting errors in the decisions due to errors in the data (disregarding situations where low data quality is favoring one politically influential group over another and low data quality is therefore politically desirable).

If geographic data is used for purposes it was not originally intended, for example using administrative data for environmental planning, the particulars of the quality of the data for this decision must be considered carefully.

In this contribution I have tried to show the effects of observation errors and how they lead to approximation of value describing objects and result in uncertainty in the classification. This does not give a set of dimensions for data quality, as has been attempted before (Chrisman 1985; Frank 1990; Wand et al. 1996); efforts to identify dimensions of data quality seem not to avoid the correlation between different dimensions: temporal or spatial resolution cannot be separated in two independently observable dimensions (and similarly for other dimensions). At the present time, I note simply that a definition of separable dimensions of data quality cannot be achieved.

References

Adams, D. (2002). The Ultimate Hitchhiker's Guide to the Galaxy, Del Rey.

Aristotle (1999). Metaphysics, Penguin Classics.

Asimov, I. (1957). Earth is Room Enough. New York, Doubleday.

Asperti, A. and G. Longo (1991). Categories, Types and Structures - An Introduction to Category Theory for the Working Computer Scientist. Cambridge, Mass., The MIT Press.

Bergson, H. (1896; reprint 1999). Matière et Mémoire. Essai sur la relation du corps et l'esprit. Paris, Les Presses Universitaires de France.

Burmeister, P. (2003). Formal Concept Analysis with ConImp: Introduction to the Basic Features. Darmstadt, Germany, TU-Darmstadt: 50.

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

Ceusters, W. and B. Smith (to appear 2006). Towards A Realism-Based Metric for Quality Assurance in Ontology Matching. FOIS, Baltimore, Maryland.

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., P. Fisher and R. Wadsworth (2004). Comparing of Expert Relations between Land Cover Datasets. ISSDQ'04, Bruck a. d. Leitha, Austria, Department of Geoinformation and Cartography.

Eckerson, W. W. (2006). "Data Warehousing Special Report: Data quality and the bottom line." Retrieved 08.08.2006, 2006, from http://www.adtmag.com/article.aspx?id=6321&page.

Fleck, L. (1935; reprint 1980). Entstehung und Entwicklung einer wissenschaftlichen Tatsache. Einführung in die Lehre vom Denkstil und Denkkollektiv. Frankfurt a. Main, Suhrkamp.

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.

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. (to appear 2005). A Case for Simple Laws. The Mystery of Capital and the New Philosophy of Scocial Reality. B. Smith, I. Ehrlich and D. Mark: 288.

Frank, A. (to appear 2006a). "Distinctions - A Common Base for a Taxonomic Calculus for Objects and Actions." Spatial Cognition and Computation.

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

Frank, A. U. (1990). Qualitative Spatial Reasoning about Cardinal Directions, University of Maine, NCGIA.

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. (2001). The Rationality of Epistemology and the Rationality of Ontology. Rationality and Irrrationality, Proceedings of the 23rd International Ludwig Wittgenstein Symposium, Kirchberg am Wechsel, August 2000. B. Smith and B. Brogaard. Vienna, Hölder-Pichler-Tempsky. 29.

Frank, A. U. (2003a). 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. (2003b). Pragmatic Information Content: How to Measure the Information in a Route Description. Perspectives on Geographic Information Science. M. Goodchild, M. Duckham and M. Worboys. London, Taylor and Francis: 47-68.

Frank, A. U. (to appear). Ontology for GIS. Vienna, Technical University Vienna, Institute for Geoinformation and Cartography.

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.

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

Gruber, T. (2005). "TagOntology - a way to agree on the semantics of tagging data." Retrieved October 29, 2005., from http://tomgruber.org/writing/tagontology-tagcapm-talk.pdf.

Gurda, R. F., D. D. Moyer, B. J. Niemann and S. J. Ventura (1987). Costs and Benefits of GIS: Problems of Comparison. International Geographic Information Systems (IGIS) Symposium (IGIS'87), Arlington, Virginia.

Heidegger, M. (1927; reprint 1993). Sein und Zeit. Tübingen, Niemeyer.

Hunter, G. J. and E. Masters (2000). What's Wrong with Data Quality Information? Abstracts. International Conference on Geographic Information Science, Savannah, GA.

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

Krantz, D. H., R. D. Luce, P. Suppes and A. Tversky (1971). Foundations of Measurement. New York, Academic Press.

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.

Leinfellner, E. (1978). Ontologie, Systemtheorie und Semantik, Duncker & Humblot GmbH.

Mac Lane, S. and G. Birkhoff (1991). Algebra Third Edition. Providence, Rhode Island, AMS Chelsea Publishing.

Marx, K. (1867; translated reprint 1992). Capital: Volume 1: A Critique of Political Economy, Pengiun Classics.

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.

Meadows, D. H., D. l. Meadows, J. Randers and W. W. Behrens III (1972). Limits to Growth. New York, Universe Books.

Mittelstraß, J. (2003). Transdisziplinarität - wissenschaftliche Zukunft und insitutionelle Wirklichkeit, Uvk.

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

North, D. C. (1981). Structure and Change in Economic History. New York, London, W W Norton &Company.

Pestel, E. (1989). Beyond the Limits to Growth: A Report to the Club of Rome. New York, Universe Books.

Raubal, M. and W. Kuhn (2004). "Ontology-Based Task Simulation." Spatial Cognition and Computation 4(1): 15-37.

Ricardo, D. (1817; reprint 1996). Principles of Political Economy and Taxation, Prometheus Books.

Robinson, V. B. and A. U. Frank (1985). About Different Kinds of Uncertainty in Collections of Spatial Data. Seventh International Symposium on Computer-Assisted Cartography, Auto-Carto 7, Washington, D.C., ASP and ACSM.

Rosch, E. (1978). Principles of Categorization. Cognition and Categorization. E. Rosch and B. B. Lloyd. Hillsdale, NJ, Erlbaum.

Sartre, J. P. (1943; translated reprint 1993). Being And Nothingness. New York, Washington Square Press.

Schneider, U., Ed. (1996). Wissensmanagement - Die Aktivierung des intellektuellen Kapitals, Frankfurter Allgemeine Zeitung GmbH.

Schopenhauer, A. (1819 & 1844; translated reprint 1966). The World As Will and Representation (Volume 1 & 2), Dover Publications.

Searle, J. R., Ed. (2001). Rationality in Action, MIT Press.

Simon, H. (1956). "Rational choice and the structure of the environment." Psychological Review 63: 129-138.

Smith, B. (1998). Basic Concepts of Formal Ontolgy. Formal Ontology in Information Systems. N. Guarino. Amsterdam Oxford Tokyo, IOS Press: 19-28.

Timpf, S. (2002). "Ontologies of Wayfinding: a Traveler's Perspective." Networks and Spatial Economics 2(1): 9-33.

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.

Vckovski, A. (1997). Interoperability and spatial information theory. International Conference and Workshop on Interoperating Geographic Systems, Santa Barbara, CA (3-6 December, 1997).

Wand, Y. and R. Y. Wang (1996). "Anchoring data quality dimensions in ontological foundations." Communications of the ACM 39(11): 86-95.

Whitehead, A. and B. Russell (1910-1913). Principia Mathematica. Cambridge, Cambridge University Press.