Extension of Metadata for Geographic Information by Process Models

Andrew U. Frank

Dept. of Geoinformation
Technical University Vienna
frank@geoinfo.tuwien.ac.at

Abstract. The database tradition asks for a static description of the data. The concept of objects has become a central idea in software engineering and has been discussed widely in the GIS literature. The object concept is difficult for geographic data, where objects can overlap and which often have undetermined boundaries like *The Foothills* or *North Sea.* The temporal identity of political entities like Poland is also open for discussion. A static description is not sufficient to properly describe geographic data to allow integration. - In this paper, the dichotomy in GIS between raster and vector is revisited from an ontological point of view. Unlike purely philosophical debates, pragmatic solutions are sought. We start with an ontology based on a continuous field model of reality and construct a secondary object ontology from it (perhaps better called an epistemology). The same reality can be subdivided into objects, especially geographic objects in multiple ways (e.g., overlapping objects). The semantics problem of interoperability can often be traced back to a difference in the way reality is subdivided into objects. The incompatibility of two datasets, one based on ownership, the other based on occupancy parcels, is a well-known example. To achieve interoperability, the connection from each kind of object to the common (continuous) reality must be reconstructed. These connections are geographic processes, which must be included in the metadata description to make them semantically more complete. It is possible to use the result from this analysis as the base to construct modeling languages, for example, for databases. To make the approach practical, improved data quality measures must be associated with each data element.

1 Introduction

Semantic interoperability (often) requires integration of models of reality. Human beings maintain models of the world in many forms – we use complex models to organize our daily life and simple models in administration – e.g., models of relations of individuals with a bank corporation. The construction of Artificial Intelligence applications or the design of databases require the formalization of models of the information about the world humans collect and help them to act intelligently (McCarthy and Hayes 1969). The models constructed reflect in some form the concepts the designer maintains himself of reality, which is expected to be shared among the future users of the information systems. However, often designers of GIS have views that differ from the user's view – which creates difficulties for the user to understand and use the system (Campari 1992; Campari 1993; Campari, Paterno et al. 1993; Campari and Frank 1995; Montello 1995). If a GIS includes several designers, then the differences in their view of the world create the so-called 'interoperability' problem (Buehler and McKee 1996; OGC 1996; Tryfona and Sharma 1996; Goodchild, Egenhofer et al. 1997; Frank and Raubal submitted).

Many think that the way reality exists is obvious and does not require further investigation. Aside from the fact that human behavior confirms that there exists a single outside reality, the way things exist and we acquire knowledge about them, is far from simple. The philosophical discipline of ontology has battled with the problem at least since Aristotle's Metaphysics. It is doubtful if this can ever be solved satisfactorily, but there are working, partial solutions, as demonstrated by many existing systems. The examples studied here show the use of a continuous model of reality and processes linking different object types. The commonality of such solutions can be generalized and exploited, for example, to design generic solutions to the semantic interoperability problem.

In this paper I want to connect models of spatio-temporal reality by models of processes. The ontological and epistemological viewpoint describes the methods used to conceptualize (i.e., understand) the world. If multiple models are used, then we have to connect them by the processes that relate them. The contribution to interoperability is the clarification of the connection between different models. Interoperability of two systems requires an integration of the models of reality used for each – this is the so-called semantics problem. If two models must be connected to interoperate, we must – explicitly or implicitly – establish formal connections between the two models. For geographic data, this connection is most often given through the common reference to space and spatial processes.

The problem of semantic interoperability is much more important for GIS than it is for other databases. The variability of modeling reality in the geographic realm looms larger than, say, in the administrative one. An administrative situation is based on the conceptual fundament of civil law; the objects recognized in each legal system are fixed and there is an efficient strategy to resolve differences in interoperation (the courts). Nothing similar exists for geography. Boundaries of undetermined spatial objects are fixed at the users' convenience (Timpf, Burrough et al. 1994; Burrough 1996; Burrough and Frank 1996). Indeed, it seems that one of the fundamental properties of space is the potential to support various and contradicting subdivisions in objects.

Unlike administrative database, spatial objects in GIS are linked to space and this provides an automatic and inherent linkage between the objects. To exploit this commonality, we propose to include process models linking datasets from different sources and describing different aspects as part of the metadata.

The paper describes first an ontological model - a model of reality as it exists – and separates from this other models that should be better described as epistemological (these models are not intended as efficient implementations, but stress conceptual aspects). It then sketches a data model for spatial-temporal data. Two case studies show the idea of integration of data based on common reference to space and explain how models of processes are used. From this we conclude that an extension of metadata for GIS with process models is a step on the way to the semantic interoperability.

2 Modeling Tools

This paper concentrates on the tools we use to model reality – not about actual models of reality. It discusses the conceptual tools an analyst has at his disposition to describe reality. Discussion of modeling tools – in the database community called *data models* – is a topic discussed in computer science and artificial intelligence since its inception. A landmark conference discussed in 1984 the different approaches the database, artificial intelligence and programming language research community took (Brodie, Mylopoulos et al. 1984). A new, integrating approach is UML (Rational 1997; Eriksson and Penker 1998).

A spatio-temporal database, indeed any database builds a model of reality. If this data model used is closer to the conceptual or cognitive models humans use, it is easier for the designer to produce an appropriate database schema. The translation of his view of reality to a formal description is simpler and requires fewer steps. It is likely that the model contains fewer errors. If the modeling language is closer to a computer implementation, constructing the database and achieving acceptable performance is likely guaranteed. In the past, modeling tools have been more influenced by implementation consideration (the object model in C++ (Stroustrup 1986) is perhaps the most recent and extreme example). Models that are close to implementation are easier to formalize and implement, but they make the task of the analyst difficult. In the following the stress is on conceptual tools and not on implementation.

In a complex system – and Geographic Information Systems is certainly one of them – multiple modeling languages are used and must be logically connected. The more complex a model, the more likely it is in conflict with another model. This is recognized as the ‘impedance mismatch’ of relational database with ordinary programming language: relational databases handle relations, which databases have to process as a collection of tuples. A similar problem is caused now by the differences between object models used in databases or programming languages.

3 Ontology or Ontologies?

In this section the important difference between a model for an ontology and a model for epistemologies will be discussed and different levels of ontology sketched. Already in 1977 McCarthy pointed out this important difference (McCarthy 1977); unfortunately, the community did not follow this differentiation.

Briefly, ontology is the science of what is and epistemology is the science of how we describe the world. Philosophers have battled with these questions for several thousand years and have created slightly varying interpretations, debating fine points, but not finding satisfactory general answers.

Artificial intelligence has then picked up the concept of ontology and given it its own pragmatic meaning (which others and I have then carried over to the GIS field). I want here to reconsider the fundamental meaning and propose pragmatic definitions, which avoid some of the inconsistencies in the use of the terms. For example, if ontology describes what is real and we accept that there is only one reality, one must conclude that there cannot be multiple ontologies - as suggested in *Spatial Ontology: A Geographical Information Point of View* (Frank 1997).

The difficulty with multiple ontologies becomes only a visible problem if databases with differing ontologies are linked together – the semantic interoperability problem. A review becomes necessary and a finer terminology required. To this we provide the base in this section, which also indicates the path towards a solution.

The discussion of the description (model) of reality must differentiate several levels: there is the level of the reality and its detailed description (in database parlance 'instances'); there is the level of the description of the types of objects in the language of a data model (in database parlance 'schema information' or metadata); and then there is the level of the data model (which would be meta-metadata). The differentiation between ontology and epistemology applies on each level.

3.1 Ontology

Ontology is the theory of being, the discussion of what exists, independent of a human observer. As long as we accept that only a single reality exists, then there is also only a single ontology.

Discussion of ontology often concentrates on a classification of objects that are somewhat similar or exist in a similar way. Aristotle differentiates substantial (physical) objects and 'accidents' that are related to a physical object (e.g., a headache). Smith has discussed geographic objects in this context (Smith to appear). In artificial intelligence discussions, the ontology section often lists the types and subtypes of objects that are relevant for the problem at hand – the ontology section defines the *universe of discourse* and lists the things that exist.

3.2 Epistemology

The term epistemology is less often used than ontology; it denotes the system of descriptors for the concepts of our understanding of reality. I propose to use it in contrast to ontology, which describes reality independent of human interaction or perception, whereas epistemology describes the knowledge (believes) cognizant entities, e.g., humans construct of this reality. One notes immediately that, strictly speaking, our discussion is always restricted to epistemology. The reality cannot be discussed – only ideas about it.

3.3 Confusion between Ontology and Epistemology

The difference between ontology and epistemology is extremely important for the modeling done in data processing systems. Many systems are intended to model reality, but the data in the system are the result of human cognitive processes and thus represent the view a human has of reality. It does not surprise then that different persons, collecting data of the same reality, come to different perceptions and represent their impressions differently. The database, which is intended as an objective representation of reality, becomes a model of the subjective perception of the human collecting the data. The descriptions resemble more epistemologies than ontologies. Well known and often discussed are practical experiments with classification of biotope from airial photographs, the classification of wetlands or the construction of soil maps: different experts produce surprisingly different models. When experts are asked to produce an objective model of reality – an ontology – they collect and code their subjective impression, an epistemology.

This leads to unproductive discussions – a user taking the model as an ontological model will not understand the possibility for alternative representations. After all, there is only one reality and one ontology! Linking such data collections shows differences due to different encodings; one then often assumes that these are errors and tries to determine which solution is correct.

4 Model of Reality

All human activity demonstrates that a single reality exists outside and independent of the observers. What are the maximum characteristics we can give of this reality independent of humans? What exists, even if no human observes it?

4.1 Continuous Model of Reality

An often-used model of reality, especially in physical sciences, is continuous. For every point in space and time, observations of properties can be made (Goodchild 1992). I assume here, without philosophical and physical investigation, that all physical properties of the world can be explained in a continuous model. Physical laws are described as differential equations (this is probably in principle the case for all physical laws).

Such a situation model is, in a nutshell, a function, which gives for any point in space and time and for any property name a value

$$F (x, y, z, t, a) = v \qquad (\mathbf{1})$$

For this model, we assume a three-dimensional space, continuous in all three dimensions. Reality evolves along a single time dimension - for each point in time a state of reality can be given. Time is ordered such that causality can influence only later situations. Physical methods exist to observe the properties at a point in space and time. Different methods to measure may lead to different values for a property, but in principle these can be related to each other.

I exclude here the extremely large and the extremely small (i.e., relativistic effects and Heisenberg's law of indeterminism); they are not relevant for the macroscopic world, which humans normally perceive and which is the subject of Geographic Information Systems.

4.2 Cellular, Discrete Computer Models of Reality

Computerized models must be discretized to be mapped to a finite representation, which can be handled in a computer.

Cellular models are characterized by a regular subdivision (discretization) of the spatio-temporal world in cells, where each cell is described with a set of property values. Discretization is here, without loss of generality, represented by the use of the representable subset of the whole numbers (e.g., the type *Int*); the use of floating point numbers (again the representable subset) is equivalent, but causes some difficulties due to the unevenness of distribution of representable numbers.

$$F\ (x, y, z, t, a) = v \qquad \text{with } x, y, z, t \text{ element of } Int \tag{2}$$

Current computers impose restrictions on the size and complexity of cellular automata, and it is impossible, in principle, to construct complete representations of reality; models are always limited and must concentrate on the aspects relevant for the task.

5 Dominance of Objects in Epistemology

In order to make sense of the physical world, where – in principle – everything is linked to everything, humans structure their perception of the environment to form grouped objects (Frank 1995). From the very large number of relations between objects people try to identify those that remain stable for longer periods of time and therefore help to model the world in a more economical representation than the interaction of spatial atoms (for a humorous account on alternative object forming see (McCarthy 1977)). This is an application of a principle of 'economy of thought' (Mach). The philosophers have always assumed that objects 'really exist' (are part of the ontology), but have problems with the obvious discrepancies how objects can be formed.

It is advisable to concentrate first on objects in the physical world and exclude difficult cases. The object concept is fundamental for human thinking and appears central in linguistics (the noun), cognition. Prototypical objects are stones, fruits and people; our analysis should start with these cases and investigate the difficult cases later (e.g., mereology (Simons 1987)).

Objects are constructed as a pragmatic shortcut to reduce the amount of effort to make sense of the environment. The grouping of phenomena to objects is thus subjective and depends on the purpose. As human interaction with the environment is restricted by our physical body and the senses, and human capabilities are by and large about the same, the same interaction patterns occur and the same types of objects are appropriate. Thus a large number of conventional objects are undisputed and appear ‘objective’; they are often included in ‘ontologies’. Other groupings of phenomena to objects are possible, actually occur and cause one kind of the semantic interoperability problem. Objects are intricately linked to operations that can be performed on them or by which they are formed.

5.1 Object Identity

Objects exist and have an identity. A person can change his name, but that does not change his identity. For example, a woman remains the same person, despite the name change by marriage (and some countries then even change the social security number).

5.2 Objects in Time

The key issue in a spatio-temporal model of reality is the concept of objects and the identity of objects through time. The Greeks have pointed out that you can never bathe twice in the same river if you consider the water. To make sense of sentences like “I have been swimming often in the Danube”, we seem to conventionally assume that there is an object ‘Danube river’ with a persistent identity.

5.3 Objects in Space

Physical objects populate space, cover some area and have a physical existence (e.g., a pencil, a car). Other objects have a location in space and often even an extension, but do not have a ‘body’ (e.g., a country). Typically objects can change their location.

5.4 Property Values Describe Objects

Objects have properties, which may vary with time. This may be thought as an observation function, which returns for each object and time the value of the property of the object. It is implemented as a function from the object identifier to the attribute value.

5.5 Relations between Objects

Objects may enter relations: a picture is nailed to the wall, an apple is in the bowl on the table. Relations can change in time. This can be represented as pairs of object-identifiers, which denote the existence of particular relations.

5.6 Generality of this Model

This model can be mapped to other models. It is compatible with the relational data model and includes the concepts proposed by Codd for its extension (Codd 1979). It is quite similar to the Entity Relationship model (Chen 1976), but it can also be seen as a model in first-order predicate calculus, especially situation calculus (McCarthy 1986).

6 Resolve Semantic Interoperability by Connection to Continuous Model of Reality

6.1 Two Example Cases

With two simplistic cases we demonstrate next the connection between datasets that represent spatial objects and appear incompatible due to different object groupings. The connection is through the 'continuous model' of reality, where a causal (or probable) link between the two data sets is constructed. Such methods are widely used on an informal base and the contribution here is to construct a single framework in which this can be formalized, generalized and later implemented.

Integration of Soils Map and Land Cover. Consider the problem of integrating a soils map and a land cover map of the same area. The objects are areas of uniform soil properties and land cover by plant habitat respectively. We concentrate on the semantics issue and assume that the boundaries are given in the same coordinate system and with a known RMS for boundary points. The overlay task is to determine the areas with a given land cover and soils type (e.g., give all areas where wheat is growing on loam). A straightforward computation without integration of the two layers results in a large number of artifacts due to inaccuracies of the object formation (so-called slivers and gaps (Burrough 1986; Chrisman 1997)). This can be avoided in most algorithms by setting a tolerance for slivers, such that slivers smaller than the tolerance are removed.

This approach is justified if one assumes a continuous reality with properties for soil parameters and climatic values, etc., for each point. Plant cover is strongly influenced by the soil properties, therefore changes in soil properties will often lead to changes in the land cover; observed soil type boundaries and land cover boundaries within the observation tolerance of the boundaries are likely two distinct observations of the same boundary. It is therefore justified to merge the two (taking into account the respective RMS) and to form spatial objects with uniform land cover and soil properties.

The connection between the two object layers is through a model that assumes a continuous reality and links the two phenomena observed to common causes by a process. The connection can be further employed to establish (local) conditional connection values, which give the probability to find land cover x on soil type y is v percent. They can be used as a predictor for missing values in one or the other data set or to identify values that are very unlikely and should be field checked.

Land Occupancy and Ownership. A similar problem is the integration of occupancy and ownership data. The connecting process is the right of the legal owner to control the use of his land, including who occupies it and what use is made of it. In the other direction undisturbed occupation as an owner leads to legal ownership in most countries (typically after 20 to 30 years).

Ownership boundaries could also be used to improve land use boundaries, as there is a causal link between land use and ownership: the owner controls the land use and land use changes are most likely at the ownership boundary.

6.2 Geographic Objects with Undetermined Boundaries

Approximated boundaries of geographic objects can be determined from the combination of layers, which contain data describing phenomena, which contribute to the composite property of the geographic object. For example, the object mountain contains element of 'close to high peak', 'inclination', certain land uses, etc. Computational combination of such data can be used to determine, for example, fuzzy boundaries of mountains in a region. For other geographic objects, causal connection based on processes for linking some properties can be constructed to give their approximated boundaries.

7 Formalization

The concept can be formalized. Processes are modeled as function linking property values. It can be used at the metadata and at the data level.

At the metadata level, models of spatial processes are constructed in a qualitative way, expressing influences between properties: In the simplest case, a connection is established between properties of a single element of space: soil type x leads to land use y (in a certain environment) and qualified by probability (this is Tomlin's focal operation (Tomlin 1990)). But more complex formulae, capturing more aspects of the spatial process are possible, taking into account more input data.

If models populated with data are constructed, then probabilities for the effect the value x of a property has on the value for another property can be computed and used to predict missing values or identify values that are implausible and should be checked.

The formalization consists of two steps:

- Process models based on differential equations (and formulated as difference equations) in space and time;
- Models to extract property values from the data sets for given points in space and time (these are essentially interpolation models). This is similar to the 'virtual data set concept' (Vckovski and Bucher 1998).

8 Limitations: Data Quality

The connections of one set of objects to other objects through process models and the common continuous reality the data is referred to, are limited. In addition to the quality of the data given and the difficulties to determine and describe this quality, comes the partial knowledge about the processes, which increases uncertainty. The models used in the case listed are quite uncertain, a large number of other influences exist and must be assumed constant (the famous *ceteris paribus* condition in economy). The resulting data combines the uncertainty in the input data with the uncertainty in the process model. This may be sufficient for many applications, but we are currently at a loss to describe the quality and compare it to the quality required in an automated system (Frank 1998). Substantial progress in the description of data quality parameters and the assessment of quality of models of processes are necessary to make the concept explained here automatically applicable (Jeansoulin and Goodchild 1998).

9 Conclusion

A philosophical differentiation between ontology and epistemology has led to a separation of a continuous model of reality and object models. Object models of geographic objects, even different models, are linked to the single spatial reality. These linkages can be exploited to integrate data from different sources. Data are connected by causal or probabilistic links established by processes. These connections can be used to integrate data, to fill areas with high probability when data are missing or to check data for implausible combinations.

The discussion shows how data descriptions (metadata), which is most often thought as static verbal description, is extended by process models from geographic science to construct powerful models of semantics. The language to describe metadata must be extended to include notions of process.

To make the concept generalizable and practically useful, progress with the description research in this area is urgently needed.

It remains an open question how this kind of semantic interoperability problem is differentiated from other ones. A classification of semantic interoperability issues remains to be produced.

Acknowledgements

I appreciate the efforts of Mag. Roswitha Markwart to improve this text.

References

Brodie, M.L., J. Mylopoulos, et al. (1984). On Conceptual Modelling: Perspectives from Artificial Intelligence, Databases, and Programming Languages, Springer-Verlag.

Buehler, K. and L. McKee, Eds. (1996). The OpenGIS Guide - An Introduction to Interoperable Geoprocessing. Wayland, MA, The OGIS Project Technical Committee of the Open GIS Consortium.

Burrough, P.A. (1986). Principles of Geographical Information Systems for Land Resource Assessment. Oxford, Oxford University Press.

Burrough, P.A. (1996). Natural Objects with Indeterminate Boundaries. Geographic Objects with Indeterminate Boundaries. P.A. Burrough and A.U. Frank. London, Taylor & Francis: 3-28.

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

Campari, I. (1992). Human Impacts on Coastal Regions: An Integrated Conceptual Framework. EGIS '92, Munich, FRG, EGIS Foundation, 1:791-807.

Campari, I. (1993). Cultural Differences in GIS. A Basic Approach. EGIS'93, Genoa, EGIS Foundation, 1:10-19.

Campari, I. and A.U. Frank (1995). Cultural Differences and Cultural Aspects in GIS. Cognitive Aspects of Human-Computer Interaction for Geographic Information Systems. T. L. Nyerges, D. M. Mark, R. Laurini and M. Egenhofer. Dordrecht, Kluwer. 83: 249-266.

Campari, I., F. Paterno, et al. (1993). The Design and Specification of a Visual Language: An Example for Geographic Information Systems Application. EGUK'93, EGUK Foundation.

Chen, P. P.-S. (1976). "The Entity-Relationship Model - Toward a Unified View of Data." ACM Transactions on Database Systems 1(1): 9-36.

Chrisman, N. (1997). Exploring Geographic Information Systems. New York, John Wiley.

Codd, E. (1979). "Extending the database relational model to capture more meaning." ACM TODS 4(4): 379-434.

Eriksson, H.-P. and M. Penker (1998). UML Toolkit. New York, John Wiley.

Frank, A. and M. Raubal (submitted). "Formal Specifications of Image Schemata - A Step to Interoperability in Geographic Information Systems." Spatial Cognition and Computation.

Frank, A.U. (1995). The Prevalence of Objects with Sharp Boundaries in GIS. Geographic Objects with Indeterminate Boundaries. P. A. Burrough and A. U. Frank. London, Taylor & Francis. GISDATA Series II, 29-40.

Frank, A.U. (1998). Metamodels for Data Quality Description. Data Quality in Geographic Information: from Error to Uncertainty. R. Jeansoulin and M.F. Goodchild. Paris, Editions Hermès.

Frank, A.U. (1997). Spatial Ontology: A Geographical Point of View. Spatial and Temporal Reasoning. O. Stock. Dordrecht, Kluwer Academic Publishers.

Goodchild, M.F., M. Egenhofer, et al. (1998). Conclusion Report of Interop'97. NCGIA.

Goodchild, M.F. (1992). "Geographical Data Modeling." Computers and Geosciences 18(4): 401-408.

Jeansoulin, R. and M.F. Goodchild, Eds. (1998). Data Quality in Geographic Information: from Error to Uncertainty. Paris, Editions Hermès.

McCarthy, J. (1977). Epistemological Problems of Artificial Intelligence. IJCAI-77, Cambridge, MA.

McCarthy, J. (1986). "Applications of Circumscription to Formalizing Common Sense Knowledge." Artificial Intelligence 28: 89-116.

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.

Montello, D.R. (1995). How Significant are Cultural Differences in Spatial Cognition? Spatial Information Theory - A Theoretical Basis for GIS (Int. Conference COSIT'95). A.U. Frank and W. Kuhn. Berlin, Springer-Verlag. 988: 485-500.

OGC (1996). The OpenGIS Abstract Specification: An Object Model for Interoperable Geoprocessing. Wayland, MA, Open GIS Consortium.

Rational (1997). UML Notation Guide, Rational Software Corp.

Simons, P. (1987). Parts - A Study in Ontology. Oxford, Clarendon Press.

Smith, B. (to appear) Objects and their environments: from Aristotle to ecological ontology. Life and Motion of Socio-Economic Units. A.U. Frank, J. Raper and J.-P. Cheylan. London, Taylor & Francis.

Stroustrup, B. (1986). The C++ Programming Language. Reading MA, Addison-Wesley.

Timpf, S., P. Burrough, et al. (1994). Concepts and paradigms for spatial data: geographical objects with indeterminate boundaries, Specialist Meeting Report, ESF Newsletter No. 4.

Tomlin, C.D. (1990). Geographic Information Systems and Cartographic Modeling. New York, Prentice Hall.

Tryfona, N. and J. Sharma (1996). On Information Modeling to Support Interoperable Spatial Databases. 8th Int. Conference CAiSE'96, in Heraklion, Greece. P. Constantopoulos, J. Mylopoulos and Y. Vassiliou. Berlin, Springer-Verlag: 210-221.

Vckovski, A. and F. Bucher (1998). Virtual Data Sets - Smart Data for Environmental Applications.