A Case Study for Semantic Translation of the Water Framework Directive and a Topographic Database

Angela Schwering[1,2], Glen Hart[1 +]

[1] Ordnance Survey of Great Britain, Southampton, U.K.
angela.schwering@uni-muenster.de
[2] Institute for Geoinformatics, University of Muenster, Muenster, Germany
glen.hart@ordnancesurvey.co.uk

SUMMARY

On the basis of a case study this paper develops a mechanism for access to multiple databases on a semantical level and outlines semantic commonalities and differences between the concepts of information communities. Although semantic translation won't overcome all differences between ontologies due to domain differences, it nevertheless serves as a bridge between the different world views, translates the meaning from one information community to another and enables portability of ontologies. In particular the paper describes the different structure of location in the conceptualizations of the datasets in the case study. The task in the scenario can be solved by integrating the world views of two particular datasets, but not simple data integration.

KEYWORDS: *Semantic Translation, Semantic Interoperability*

INTRODUCTION

Increasingly government and commerce requires complete access to available information to efficiently meet the demands placed upon them by citizens and customers. Typically this information is both heterogeneous and distributed over many databases. The solution to many tasks requires access to multiple datasets and therefore a seamless and flexible information integration crossing organizational borders.

To reuse and share geo-data successfully, integration has to be realised on not just a syntactical level but also on a semantical level. To ensure that the integration is reasonable and the interaction between datasets makes sense a formal description of the semantics and a concept of their world views is required. On the basis of a case study this paper develops a mechanism for access to multiple databases on a semantical level and outlines semantic commonalities and differences between the concepts of information communities. The task in the case study can be solved by integrating the world views of two particular datasets, but not simple data integration.

SEMANTIC INTEGRATION OF GEO-DATA

Since data is made available by providers from different information communities and is gathered from different point of views its structure might differ quite a lot. For example the properties of a river from an ecological viewpoint are very different from a river's properties from a transport viewpoint. In the first case it may be used for drinking water and in the second it is used for routing.

The context and the domain in which knowledge is used has a great influence on the meaning of terms and therefore also on the modelling. Each information community has its own world view - called the conceptualization of the world. Studer defines a conceptualization as "an abstract model of some phenomenon in the world by having identified the relevant concepts of that phenomenon" (Studer, 1997).

Within ontologies the semantic of these relevant concepts are formalized. An ontology is an explicit specification of a conceptualization to which the vocabulary of one information community refers (Gruber, 1993; Gruber, 1995). Ontologies support the integration task by representing the semantic of the data explicitly and therefore making the data semantically enabled.

The specific task of the following case study is dependent on multiple data sources derived from two or more domains (see fig. 1). The ontologies representing each conceptualization of these domains overlap describing the same real world objects, but don't necessarily agree on a shared conceptualization of these real world objects. Due to different world views, the conceptualizations of the real world objects differ in the ontologies of both information communities. To overcome these differences all participants of a communication process have to agree on the terms, their meanings, relations and constraints: the ontological commitment (Studer, 1997). Each information community commits to its own conceptualization, but to enable semantic translation, some common, more abstract, shared conceptualization is needed, which describes the semantic space with a reduced complexity. This basic ontology will overcome the semantic differences such that interaction between ontologies becomes possible to combine data from the data sources and enable matching or translation of concepts between different information communities. In the following the differences and commonalities of the two conceptualizations are analyzed to find a way for interaction between these ontologies to combine data from the data sources to meet the needs of the task. Although semantic translation will not be able to overcome all differences between the ontologies due to domain differences, it nevertheless serves as a bridge between the different world views, translating meaning from one information community to another and enabling portability of ontologies.

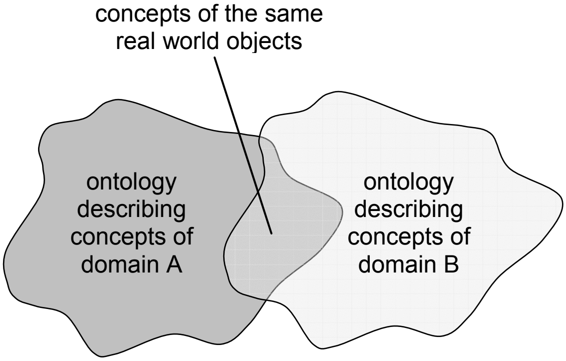
concepts of the same real world objects
ontology describing concepts of domain A
ontology describing concepts of domain B

Figure 1: Overlap of ontologies describing different domains

To formalize definitions of terms in the ontologies a variety of languages can be used. In our case study simple graphical figures are used as language to formalize terms, their meaning and relations, because we want to abstract from restrictions caused by representation languages like OWL and also from inter-language translation problems.

THE CASE STUDY

The task underlying the scenario is to improve the ecological status of a water body, e.g. a river. For this purpose rivers with bad ecological status are identified. Secondly from the information about the topological relationships between the water body and other features conclusions can be drawn as to what might have caused the bad ecological status of the river since the quality of the water will be dependant on the influx of material from the surrounding land. To solve this problem different kinds of datasets are needed. Information defining ecological status and information pertaining to the status of a particular river can be found in the Water Framework Directive (WFD) (European Parliament, 2000; Vogt, 2002) and within the databases of the relevant environment agency. Information of topological relationships of features is saved in a topographical database, e.g. from a mapping agency (e.g. Ordnance Survey, 2001). Through the interplay of several ontologies the relevant information can be gained and the task can be solved.

After the data requirements for the improvement process are determined, ontologies containing the relevant data are searched for. A number of different ontologies are needed to solve this task: The WFD ontology describes the meaning and the relations of the data of the WFD, the topographic ontology describes the semantics of a mapping agency's database, e.g. OS Master Map. To answer the task of the case study industrial or agricultural utilization upstream have to be searched for. An ontology is needed that describes the relations between causes of pollution and their effects. This ontology won't be discussed deeply here, since this paper focuses on the problems of semantic integration of the WFD ontology and the topographic ontology.

The environment improvement process has several steps : Firstly rivers with bad ecological status are identified - the WFD provides this information. The WFD lists all water bodies in the investigation area and distinguishes between rivers, lakes, coastal-, transitional water bodies and groundwater bodies. Fig. 2 shows the structure of the terms used in the WFD to describe water bodies.

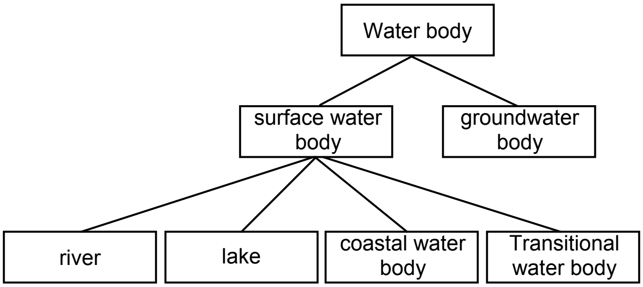

Figure 2: Terms for water bodies in the WFD (EU, 2000)

The second step is to find the information of the same real world object in a topographical database. Usually a topographical database has a very different conceptualization of water bodies and what constitutes a river.

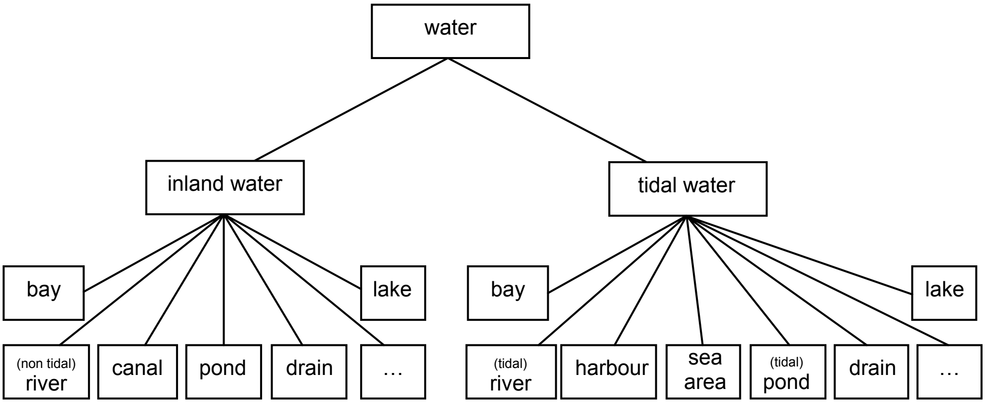

Figure 3: Terms for water in a topographic database (Ordnance Survey, 2001)

Fig. 3 gives an overview of some of the terms used in a topographic database to describe different kinds of water bodies or parts of water bodies.

SEMANTIC COMMONALITIES AND DIFFERENCES

Semantic translation is the bridge between both conceptualizations of the world. Both conceptualisations describe the same real world object and call it river, but due to domain differences they hold very different information about it.

Definition of Terms

There are semantic differences in the meaning of what a river is: the WFD defines a river as a "a body of inland water flowing for the most part on the surface of the land but which may flow underground for part of its course." (European Parliament, 2000), though the topographical database underlines the topological relations to other features: "Water flowing in a definite channel towards the sea, a lake or into another river" (Ordnance Survey, 2001). A river that flows underground is not recognised within the topographical database as it cannot be recorded or represented.

Properties of Concepts

Not only does the different meaning of the term river cause problems during the integration process, but also the purpose of the data is not the same. Both databases consider river in a different context and focus on different content and properties.

In the following we consider the characteristics of rivers from the Water Framework Directive's and the environment agency's point of views and from the viewpoint of a topographical database.

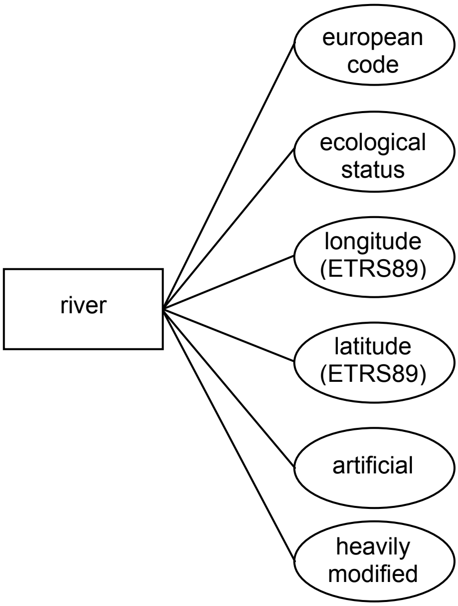

Figure 4: Real world object "river" in the WFD (EU, 2000)

This figure gives an overview about some properties of rivers in the WFD. The characteristic properties considered to be important in the WFD are very different from other databases. The WFD also contains very detailed information about the ecological status of the water body like data about the physical, chemical and the fresh water ecological status.

Each river segment is identified by one unique European code. It provides an optional attribute to specify the location of the river: it is the mathematical centre of the river described by two coordinates in the ETFS89 - longitude and latitude. But the mathematical centre itself is not well-defined, both in terms of its mathematical computation and its meaning. Furthermore describing the location of a river by a simple point is very imprecise and vague.

The environment agency provides more detailed information about the location of a river. It covers the whole river course on a very detailed level. Since the WFD is produced by the local environment agencies, usually information exists to map rivers with an European identification codes of the WFD to the ones from the environment agency and vice versa.

The WFD only describes rivers, but the topographical database distinguishes between rivers and the river mouth (tidal river) and according to the width of rivers they are called ditch, stream, river etc.

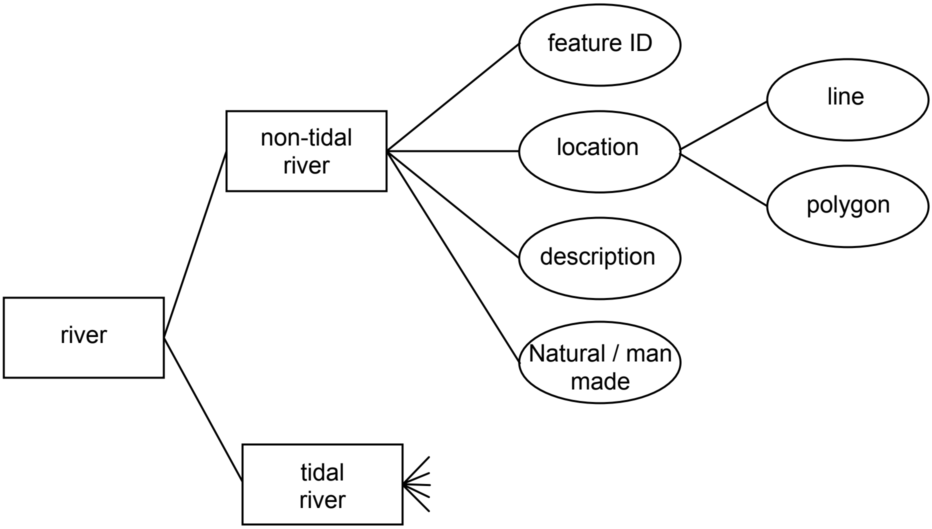

Figure 5: Real world object "river" in a topographical database (Ordnance Survey, 2001)

In the topographic database the location is represented by a simple line or a polygon, e.g. with Gauß-Krüger coordinates or the National Grid.

The WFD differentiates between artificial, heavily-modified and natural water bodies, in contrast the topographical database describes them as man-made and natural. Of course, both have very different norms to categorize the water bodies.

Context of Concepts

The WFD and datasets of the relevant government environment agency can be used for the transport network and inland navigation. Ecological aspects like drinking water supply, habitat for animals and plants, nature conservation and flood defence are important fields of application as well. Therefore datasets of an environment agency often contain information about the chemical and biological status of the water. The river course is not described just by one polygon but also by the flood water line and information about the flood plains. The river is considered as one object with tributary rivers from their sources to their mouths.

From the different purpose of usage result differences in storage of the identification code and the location of a river. The territory is divided into river basins and sub-river basins. The numbering follows the concept shown in Fig. 6.

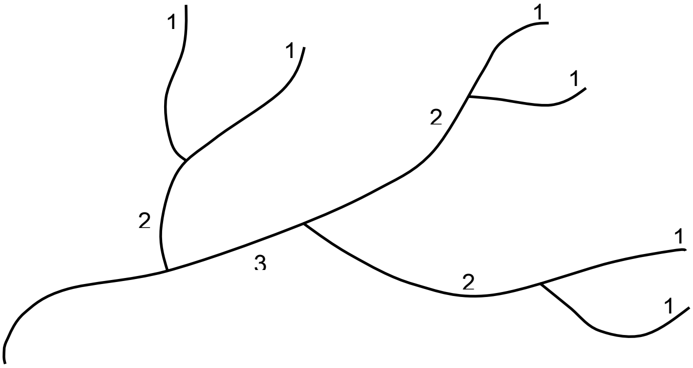

Figure 6: Identification of rivers

In most topographical databases rivers are represented by spatial disjunctive sections, with sections being divided by bridges and other structures, due to the original need to generate paper mapping. Figure 7 shows rivers divided in two fractions by a footbridge. Increasingly national mapping agencies such as Great Britain's Ordnance Survey are recognising the importance of representing geographic entities as complete features. Thus Ordnance Survey has changed its capture specification to ensure that new survey detail includes areas of rivers obscured by other features are now captured. The existing database is also being revised retrospectively to ensure that existing detail is represented in the future as continuous features through a combination of semi-automatic and manual means.

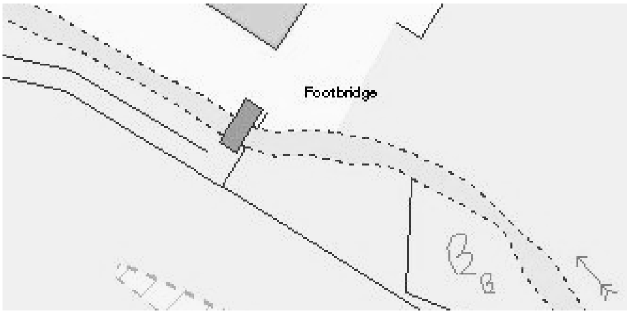

Figure 7: River features in a topographical database[49]

In the topographical database used in the case study "non-tidal rivers more than 1.0 m wide have their limits captured by depicting each bank. Non-tidal rivers less than 1.0 m wide have the centre of their alignment captured using a single line. The width to be shown is that at normal winter level. The direction of flow is indicated" (Ordnance Survey, 2001). The coordinates of the polygon or the line can be stored as Gauß-Krüger Coordinates or the National Grid. Ecological data is not captured and flood defences although many will be recorded will rarely be recognised explicitly as such.

SEMANTIC PROBLEMS IN THE INTEGRATION PROCESS

One of the main challenges during the integration process is the identification of corresponding features in the different conceptualizations and the transformation of them. Exemplary for many different semantic integration problems this paper describes the integration of different river concepts.

[49] Sample from OS Master Map, visualized with OS Master Map Viewer from Snowflake Software.

The first semantic problem occurs during the identification of the corresponding features. Though both world views contain conceptualizations of water bodies they represent very different things. The definition of a river described in the ontology gives some information what constitutes the feature river. To be processed by a computer this definition must be available in a formal way and not in plain text. This definition usually gives a general idea of the real world object.

Secondly the concept of a river is described from a different perspective and therefore they focus on other properties and implement different relations to other objects. But still there are some common basics, e.g. both conceptualizations contain information about the location, but the meaning as well as the granularity are not the same. To make sure that the identified rivers are exactly the corresponding objects, we propose to prove they have the same location, which refer to different reference systems in the different conceptualizations. But since the geographic reference systems are well defined and transformation between different reference frames exist transforming the location saved in ETFS89 to coordinates like Gauß-Krüger or the National Grid is no problem. The water framework directory defines location as the mathematical centre of the water body as the location. Transforming these coordinates to the suitable geographical reference system and checking if this point is inside the polygon saved as the location for the river in the topographical database is one method for proving the identity of both objects, but still it is a very imprecise one. Therefore it's better to use a more detailed model of location. The environment agency considers a river as a cohesive object and contains information about the course of the main river and its affluxes. The topographical database models the location of a river by saving many polygons for each section of the river, which might be disjunctive. Methods like edge matching can help to find the corresponding objects in the database of the environment agency and the topographical database.

The third problem during the integration process is the transformation of a concept of one conceptualization to another concept within a different conceptualization. This means that the whole concept - its label, properties, relations and constraints - are translated into the corresponding analogon in the other conceptualization.

CONCLUSION AND FUTURE WORK

Though datasets often describe the same real world object the data is very heterogeneous because of different conceptualization of the world. Due to different world views they consider different things to be important, but in many cases they still either

- describe the same or similar things, but give them different labels, e.g. coastal water body in the WFD and sea area in the topographic database.
- describe different things, but give them the same labels or have different conceptualizations and use the same name for it, e.g. river or location.
- describe data on different levels of granularity, e.g. while the topographic database distinguishes between lakes, lochs, reservoirs, ponds, balancing ponds, curling ponds, tidal ponds … the WFD has only one corresponding object “lake”.
- describes data which doesn’t exist in the other database at all, e.g. the WFD provides detailed information about the water quality, chemical and biological measurement of the water and fish. A topographical database doesn’t contain any of this kind of data.

This paper outlines semantic commonalities and differences between the concepts of information communities based on comparing the vocabulary used in the datasets, the definition of terms, their properties and the context within concepts are used. To enable semantic translation the differences have to be formalized and the relations of the corresponding concepts are analyzed. In the case study the semantic translation between the different domains is based on a more simple and abstract concept of a river on which both conceptualizations of a river relate. This implies that a fundamental consensus exists about the concept of a river.

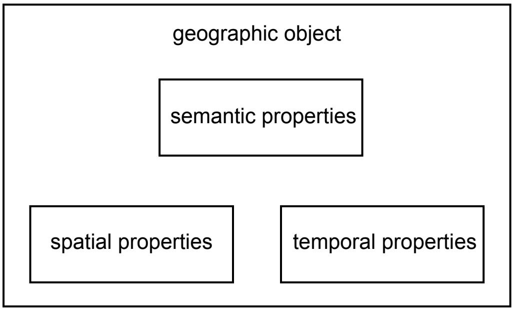

Figure 8: Aspects of semantic translation of a geographic object

Because geographic data consists of spatial, temporal and thematic data, using purely semantic data for the transformation won't enable full interoperability of different instances of the data, but the interplay of all aspects of a geographic object like spatial, temporal and semantic properties. By using a simple and more abstract concept of a river in combination with the spatial aspect of the river like the location, the semantic transformation from one spatial space to another becomes possible. In the case study the combination of the feature "river" and its "location" makes it possible to identify one real world object non-ambiguously.

The next step of the research is to identify the potential to generalize the results for other case studies and as well to analyze the relations of corresponding concepts. After identifying similar concepts the nature of this relation (synonym, hypernym, homonym, topologic relations …) is examined, because the nature of relation is very important to make it useful and use it in the correct way.

ACKNOWLEDGEMENT

We would like to thank Prof. Werner Kuhn for providing constructive comments and suggestions. The directions and ideas he gave are very valuable for our research.

REFERENCE

Gruber, T.R. (1993). A Translation Approach to Portable Ontologies, Knowledge Acquisition, Vol.5 No. 2, pp. 199-200. http://gicl.mcs.drexel.edu/people/regli/Classes/KBA /Readings/KSL-92-71.pdf

Gruber, T.R. (1995). Toward Principles for the Design of Ontologies Used for Knowledge Sharing, Int. Journal of Human-Computer Studies, Vol. 43, pp.907-928. http://citeseer.nj.nec.com/gruber93toward.html

Guarino, N.; Poli, R. (1995). Formal Ontology in Conceptual Analysis and Knowledge Representation. International Journal of Human and Computer Studies, Vol. 43 No. 5/6. http://citeseer.nj.nec.com/guarino95formal.html

Kuhn, W. (2003). Semantic Reference Systems. International Journal of Geographical Information Science, Vol. 17 No. 5, pp. 405-409. http://musil.uni-muenster.de/documents/SemanticReferenceSystems_final.pdf

Ordnance Survey (2001): Master Map real-world object catalogue. http://www.ordnancesurvey.co.uk/oswebsite/products/osmastermap/pdf/realWorldObjectCatalogue.pdf

Studer, R.; Benjamins, R.; Fensel, D. (1998). Knowledge engineering: Principles and methods. Data and Knowledge Engineering, Vol. 25, pp. 161-197. http://www.ubka.uni-karlsruhe.de/cgi-bin/psgunzip/1997/wiwi/33/33.pdf

European Parliament (2000). EU Water Framework Directive, Directive 2000/60/EC, 23/10/2000. Official Journal (OJ L 327), 22 December 2000. http://europa.eu.int/comm/environment/water/water-framework/index_en.html

Vogt, J. (2002). WFD Working Group GIS: Guidance Document on Implementing the GIS Elements of the Water Framework Directive. http://eurolandscape.jrc.it