PRACTICAL ONTOLOGY DEVELOPMENT
Some Lessons Learnt

Rolf Grütter, Bettina Waldvogel
Swiss Federal Research Institute WSL, Zürcherstrasse 111, CH-8903 Birmensdorf, Switzerland
{rolf.gruetter, bettina.waldvogel}@wsl.ch

Curdin Derungs
Department of Geography, University of Zurich-Irchel, Winterthurerstr. 190, CH-8057 Zurich, Switzerland
curdin.derungs@geo.uzh.ch

Keywords: Knowledge Representation, Ontology Engineering, Applications and Case Studies.

Abstract: Ontology engineering is a well travelled ground, at least from a theoretical point of view. Long before the Semantic Web became popular, principles for the design of ontologies were established and, recently, guidelines and methods for building ontologies were published. Despite this guidance, there are issues in practical ontology development, which are not covered in the literature. This paper discusses some problems that occurred during the manual construction of an OWL application ontology and that required design decisions by the developers.

1 INTRODUCTION

Ontology engineering is a well travelled ground, at least from a theoretical point of view [Gómez-Pérez et al., 2004]. Long before the Semantic Web became popular, principles for the design of ontologies were established (Gruber, 1995) and, recently, guidelines and methods for building ontologies were published (Kovacs et al., 2006, Noy and McGuiness, 2001). Despite this guidance, there are issues in practical ontology development, which are not covered in the literature.

This paper discusses some problems that occurred during the manual construction of a medium sized (690 concepts, 50 properties, 4486 individuals), formal (description logics $\mathcal{ALCHI}(\mathcal{D})$) bilingual (German/French) application ontology and that required design decisions by the developers. It shows by examples how these problems were solved. Based on the insights gained during ontology development the paper comes up with lessons learnt that may be helpful for other developers.

The paper is organized as follows: In Section 2 design decisions that were taken prior to the development process are described. Section 3 reports on design decisions that were taken during the development process and which were not anticipated. Section 4 discusses the described design decisions and Section 5 concludes with a list of lessons learnt.

2 A PRIORI DESIGN DECISIONS

A first design decision is related to the *scope* of the ontology: What kind of knowledge shall be represented? This decision depends on what the ontology is intended for and on the conceptualization underlying the data to be described.

In our case the ontology was intended to support non-expert users in retrieving information from a large database for national species management by expanding queries semantically. Together with former and future users of the system, we compiled a number of use cases. Documents and data were analyzed for the concepts that were assumed as (implicit) models of the respective domains by their authors as well as for the names of these concepts. Based on the use cases and on the results of the analysis, significant concept hierarchies were build for authorities, legal acts, regions, inventories (i.e. kinds of data collections), habitats, animal species, plant species, political processes, legal entities and documents.

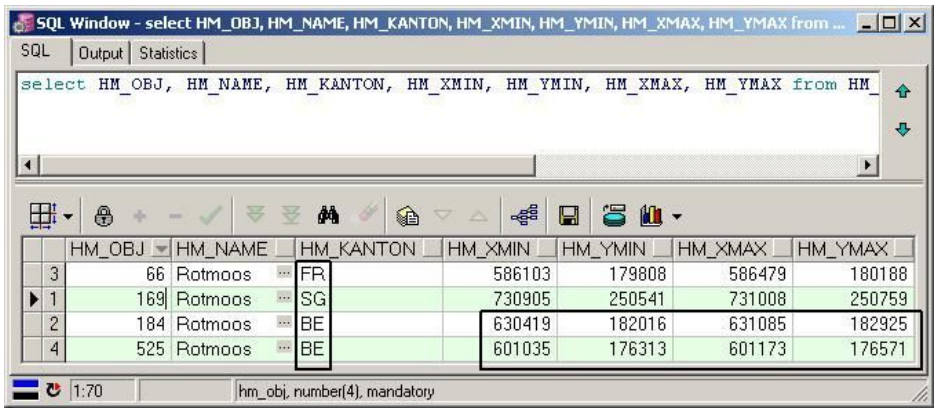

Figure 1: Entities that share the same name are differentiated based on their spatial references.

A further design decision is related to the question whether a *formal* or a *non-formal* ontology should be build and, if formal, which framework should be used. We decided to build a formal onto-logy using Web Ontology Language (OWL) (Patel-Schneider et al., 2004). The decision of adhering to the standard technique was taken in order to make the ontology (or parts thereof) reusable and also to take advantage of existing editors and reasoners to process the ontology (note that our ontology can be obtained from the authors for research purposes).

Still a further design decision is whether *individuals* should be asserted in the ontology. From the use cases we learnt that users are not only searching for (sets of) individuals (which can be retrieved by checking the satisfiability of concepts) but also for specific properties of a certain individual (in the case of a spatial object, its geometry, for instance). In order to support this kind of search we decided to also assert individuals in the ontology.

3 AD HOC DESIGN DECISIONS

Most design decisions during ontology development were related to the disambiguation of entity names and to figure out their meaning in terms of concept membership. The extent of this challenge depended on the kinds of entities to be integrated.

3.1 Is it a Name or a Taxon?

Animal and plant species in our data collections are named by well-defined Latin taxa. For these a straightforward approach could be applied: For one entity in the database one named individual was asserted in the ontology. Still the choice of the taxonomy required a decision. Domain experts do not work with a single taxonomy; they choose the most appropriate, depending on the project at hand. Ontological design, therefore, has the choice of either concentrating on a single taxonomy or considering competing classifications.

We chose a pragmatic approach and decided in favour of a single taxonomy, namely the taxonomy which is most commonly used by the domain experts working with the data.

3.2 What Kind is this Entity of?

In our data collection there are entities whose names mean different things, depending on the context in which they are used. Consider, for instance, the name “Kanton Bern” (i.e. Canton of Berne), usually as a shortcut “BE”. “Kanton Bern” can be the name of a regional authority governing an administrative unit or it can be the name of the administrative unit governed by the regional authority.

In order to figure out the meaning of “Kanton Bern” and to record it in terms of concept membership in the ontology, we looked at how it is used in documents and data in the database. “Kanton Bern” is usually used to denote the regional autho-rity. Accordingly, we asserted an individual Kanton_Bern for a concept Kanton which is subsumed by a concept Gebietskörperschaft (i.e. regional authority).

3.3 Is this the Same Entity?

In databases, entities are identified by keys, in ontologies individuals are identified by names. Entities in databases may also have names but these are not necessarily unique. Consider the name “Rotmoos”, for instance: Quite a number of entities in our data collection share this name.

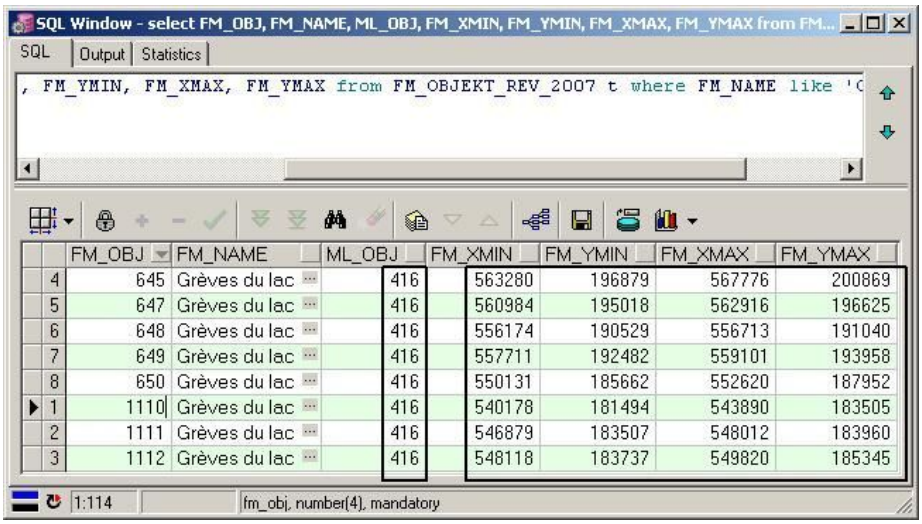

Figure 2: Entities with different geometries share the same additional type.

In order to decide whether these entities refer to the same object or not, the procedure described in Section 3.2 does not work. All are of the same (general) kind, that is, they are members of concepts that are subsumed by the concept Lebensraum (i.e. habitat). Instead, we took advantage of a special feature of these entities: They all have a spatial reference, both in terms of the administrative unit they belong to and in terms of a geometry (namely, the coordinates of a bounding box).

Using the administrative information, two of the entities sharing the name "Rotmoos" could be identified as different objects: They belong to the cantons Fribourg (FR) and St. Gall (SG) (Figure 1). In order to decide whether the two Bernese entities "Rotmoos" refer to the same object or not, the administrative information was not helpful. The entities could be identified as different objects only by comparing their geometries (Figure 1).

3.4 Is this a Different Entity?

There are eight entities with the name "Grèves du lac" in the collection of fens, whose geometries are different. According to the procedure described in Section 3.3, each of them would have to be asserted as a uniquely named individual in the ontology, for instance as Grèves_du_lac_I, Grèves_du_lac_II, … Grèves_du_lac_VIII. Would this be the right decision? A closer look at the data in the database denies the question: All eight entities correspond to a single entity in the collection of moorlands (Figure 2). If each entity was asserted as a named individual in the ontology, which one would be the individual that is both a fen and a moorland?

The solution we chose was to assert for all eight equally named entities of the collection of fens a single uniquely named individual Grèves_du_lac in the ontology and to type it as both a Fen and a Moorland.

4 DISCUSSION

4.1 A Priori Design Decisions

As mentioned in Section 2 the design of an application ontology depends on its purpose and on the conceptualization of the people who collected the data and of those who are going to use the ontology. The *vocabulary* provided by an application ontology further depends on the mother tongue of these people. Because of these dependencies, it is very unlikely that an application ontology can be reused in a different context.

The advantage of formal ontologies, such as description logics ontologies, is that they allow automatic consistency checking and reasoning. By computing entailments, new knowledge is being inferred from existing knowledge. However, formal ontologies impose rigid restrictions on the structure of the knowledge they represent. A part of the knowledge expressible in natural language cannot directly be modelled by such a formal framework: vague concepts, fuzzy information, general rules with exceptions. Whether the gain in knowledge earned by sound and complete inference procedures overweighs the loss of knowledge taken by the rigid framework is an open question.

4.2 Ad Hoc Design Decisions

As mentioned in Section 3.2 an entity name can mean different things, depending on the context in which it is used. There are different ways of dealing with context. One way is to deal with it in terms of the discourse that takes place (Kamlah and Lorenzen, 1967). A *discourse* makes use of a vocabulary, which can be specified by an ontology. Still, a vocabulary does not make up a discourse. Philosophy of science explains this by differentiating between the communicative role and the representative role of language (Kromrey, 2002). According to this differentiation, the language in which a discourse is expressed is different from the language used to represent the vocabulary. Since the design of a communication language is outside the scope of ontology engineering, the context of the discourse does not directly affect design decisions of ontology developers.

Another way of dealing with context is to consider the *situation* in which a discourse takes place (Kamlah and Lorenzen, 1967). Using a theatre metaphor, this kind of context is also referred to as the play, which is performed, together with the different scenes of that play (Laurel, 2003). What the interaction that takes place is all about, can also be referred to as a *theme*. Different from discourse, themes directly influence decisions of ontology developers. The consideration of context in ontology development is, thus, closely related to the decision on the scope of the ontology (cf. Section 2).

Differentiation of entities by analysing their geometries (cf. Section 3.3) is an established method in Geographic Information Science. The claim is that two (or more) entities are the same if they share the same (or a very similar) geometry (e.g. Sester et al., 2007). Conversely, entities sharing the same name can be differentiated according to the places they refer to. As Section 3.4 suggests the application of this method yields better results when it is combined with type information. Note that differentiation by using type information is a work-around to uncover the case where shared names point to the same broader place in the real world, which is not recorded in a database.

5 CONCLUSIONS

The main lessons learnt from the development of an application ontology as described in Section 1 are:

(i) The scope or theme of the ontology cannot be copied from the available data and pasted in the ontology; it rather has to be figured out by compiling use cases together with future users and by analyzing the data in view of the implicit conceptual assumptions that were made by their authors; (ii) entities in databases do not a priori correspond to individuals in ontologies in an enumerative way; (iii) differentiation of entities with a spatial reference by analysing their geometries does not always work.

ACKNOWLEDGEMENTS

This research has been funded by the Swiss Federal Office for the Environment (FOEN).

REFERENCES

Gómez-Pérez, A., Fernández-López M., Corcho, O., 2004. Ontological Engineering, Springer London, pp. 403.

Gruber, T. R., November 1995. Toward principles for the design of ontologies used for knowledge sharing. *International Journal of Human-Computer Studies*, Vol. 43, Issues 4-5, pp. 907-928.

Kamlah, W., Lorenzen, P., 1967. Logische Propädeutik. Vorschule des vernünftigen Redens. *Bibliographisches Institut*. Mannheim.

Kovacs, K., Dolbear, C., Hart, G., Goodwin, J., Mizen, H., June 2006. A Methodology for Building Conceptual Domain Ontologies. Ordnance Survey Research.

Kromrey, H., 2002. Empirische Sozialforschung, Modelle und Methoden der standardisierten Datenerhebung und Datenauswertung. Leske + Budrich. *Opladen, 10., vollst. überarb*. Aufl.

Laurel, B., 2003. Computers as Theatre. Addison-Wesley. Boston, 10th printing.

Noy, N. F., McGuinness, D. L., March 2001. Ontology Development 101: A Guide to Creating Your First Ontology. Stanford Knowledge Systems Laboratory Technical Report KSL-01-05 and *Stanford Medical Informatics Technical Report* SMI-2001-0880.

Patel-Schneider, P. F., Hayes, P., Horrocks, I., 2004. OWL Web Ontology Language. Semantics and Abstract Syntax. W3C Recommendation 10 February 2004.

Sester, M., von Gösseln, G., Kieler, B., 2007. Identification and adjustment of corresponding objects in data sets of different origin. In Proceedings of the 10th *AGILE International Conference on Geographic Infor-mation Science 2007*, Aalborg University, Denmark.