Designing a Bilingual Eco-Ontology for Open and Intuitive Search[1]

Bettina Bauer-Messmer and Rolf Grütter
Swiss Federal Institute for Forest, Snow and Landscape Research
Zürcherstrasse 111, CH-8903 Birmensdorf ZH, Switzerland
{bettina.bauer, rolf.gruetter}@wsl.ch

Abstract. In environmental databases there is a plethora of data, described in different terminologies, stored in different data structures and referring to distinct time periods and locations. Thus it is difficult for non-experts to find the data they are looking for. Adding an ontology-layer, which provides semantic enhancement, provides the underlying data structure for an open and intuitive search interface for non-experts. In this paper, we present the design of a bilingual eco-ontology using fuzzy mappings and loosely bridged ontologies which introduce a semantic layer in an existing environmental database storing data in German and French.

1. Introduction

Existing environmental databases hold a great variety of data, which are encoded in different languages, grouped into diverse inventories using their own terminologies and refer to distinct time periods. Our environmental database at the Swiss Federal Institute for Forest, Snow and Landscape Research in particular is characterized by the coexistence of diverse inventories with data in German and French. In order to help non-expert users with finding the data they are looking for, we propose to add an ontology-layer as the underlying data structure for an open and intuitive search. An *open* and *intuitive* search does not only look for exact matches of specified filter criteria but also searches for semantically similar database entries. It supports patterns of interaction which are familiar to the user. This kind of search is made possible with a natural language user interface.

Processing natural language input from users requires information systems, which "understand" up to a certain degree the input. The term "understand" refers to a mapping of conceptual structures in the users mind onto data structures. Ontologies are embodiments of shared conceptualizations (Gruber 1993) and therefore can semantically interpret and augment user input such as search terms.

Eco-ontologies are either defined as focusing on spatial "ecological niches" (Smith and Varzi 1999) or they stress the aspect of semantic het-

[1] Published in: Gómez, J.M., Sonnenschein, M., Müller, M., Welsch, H., Rautenstrauch, C. (eds.): Information Technologies in Environmental Engineering. ITEE 2007 – Third International ICSC Symposium. Springer-Verlag, Berlin Heidelberg (2007) 143–152

erogeneity of source communities (Fonseca et al. 2002). In this paper the term eco-ontology refers to the later definition. Eco-ontologies represent the geospatial domain, for which no consensual ontology yet exists, and as long as the ambiguities within concepts are not clarified, it will, according to Agarwal (2005), not be possible to create one.

Building a bilingual ontology for environmental data poses challenges of different types:

- **Linguistic challenge:** The conceptualizations and vocabulary for objects in the environment change over time, vary between ethnic and cultural groups and are strongly influenced by regional dialects.
- **Bilingual challenge:** The categories of terms are built based on different criteria and do not always have the same level of detail in German and French. Direct mapping between these two languages would lead to a considerable loss in expression.
- **Heterogeneous semantics challenge:** The world is ambiguous. Datasets from different communities do not share a conceptualization. Even within a single domain, e.g., biology, taxonomies can be contradictory.
- **Toponymy challenge:** Geographic locations are often known by several names. These names depend on the scale of a map (country, canton, community). Names in rural dialects differ from those used in official maps.

The work presented here is focused on the linguistic and bilingual challenges of constructing an eco-ontology which also touch the heterogeneous semantics challenge. It builds on previous work on the design and implementation of a Web-based platform for visualizing, querying and analyzing environmental data (Baltensweiler and Brändli 2004). The toponymy challenge will be addressed elsewhere (in preparation).

2. Linguistic Challenge: Defining a vocabulary and its classification

2.1 Local dialects and cultural differences

Depending on language, cultural background, social environment, education and many more factors the number of terms describing a given sector of the world varies enormously. Traditional Filipino Pygmies distinguish a wide variety of animals and plants: For example 15 breeds of bats, 75 avian breeds, 20 breeds of ants and 450 plants among many others (Levi-Strauss 1962). In general, rural dialects are very rich in terms and definitions of rural everyday life. However these dialects are far less specific in

other subject fields. The level of detail in describing concepts directly represents the importance for the individual. Thesaurus completeness is crucial for the success of information retrieval systems (Salton 1970). Therefore great care must be taken to construct a thesaurus with a level of detail which corresponds to the needs of the potential users.

2.2 Historic Classifications

There are three German terms for bush, shrub and perennial herb: Busch, Strauch, Staude. Nowadays, the term Busch is a superordinate concept of Strauch and Staude. The difference between Strauch and Staude is based on the fact, that the stems of a Strauch are wooden, and the stems of Staude are soft. This hasn't always been the case. About 150 years ago there was a distinction between Strauch and Staude based on the fact, that Staude yields (edible) fruit and Strauch does not so (Grimm 1854). Alas, there is not only a semantic shift but also the focus of attention has changed. This is reflected by the fact, that taxonomic categories are built based on different criteria, indicating that yielding fruit was an important criterion for people 150 years ago, but is no longer these days.

2.3 Orthography and spelling

The concepts of orthography and spelling emerged only recently. In 1776 Thomas Jefferson’s draft of the Declaration of Independence showed many spelling inconsistencies. Even these days the German orthography reform in 1996 lead to hot tempered discussion. Language and spelling was and still is an instrument for individual expression.

3. Bilingual Challenge: Combining conceptualizations from different languages

In the last decades multilingual information retrieval has become an ever growing research area. The architectures of these systems haven't changed much over the years (Salton 1970): thesaurus (vocabulary), translation facility (mapping functions, translation tables), hierarchical arrangement of concepts and more or less sophisticated algorithms, which represent the semantic field and context. By contrast the algorithms and data storage techniques have changed considerably since the beginnings, especially with regard to the global community participating and sharing information in the Semantic Web.

A more recent approach to multilingual information retrieval is EuroWordNet, although not being a pure ontology, it uses a top ontology to enforce uniformity and compatibility of different word-nets (Vossen 2002). Other multilingual ontologies are being built all over the world with

huge vocabularies, e.g the Chinese-English ontology or the Arabic WordNet. Furthermore efforts are taken to combine bilingual translation services in a grid.

3.1 Taxonomic Hierarchies in German and French

In German there are three terms for bush: Busch, Strauch, Staude. In French, however, there are four words: buisson, arbuste, arbrisseau and plante vivace. The definitions of these terms (disregarding the fact, that definitions vary by author) are based on different criteria. Furthermore, individual plants are categorized differently in both languages.

It is most remarkable that the classification schema is not complete and not free of overlaps. For instance plants with soft stems and a height of more than 25 cm and less than 4 m do not fit in any category. On the other hand hazelnut can be classified as both shrub and tree.

Trying to combine both German and French in a single ontology inevitably leads to inaccuracies or inconsistencies. In order to preserve the richness of detail and the historically and culturally grown categories, it is imperative to create a separate ontology per language.

In domains such as biology even in a single language several different classification schemes, so called systematics, coexist. Experts are used to reference the classification scheme whenever they report on their work.

The classification of plants shown in figure 1 is merely one of many possibilities. Plants can be classified just as well by their life-form, for instance the height above (or below) ground of the renewal bud.

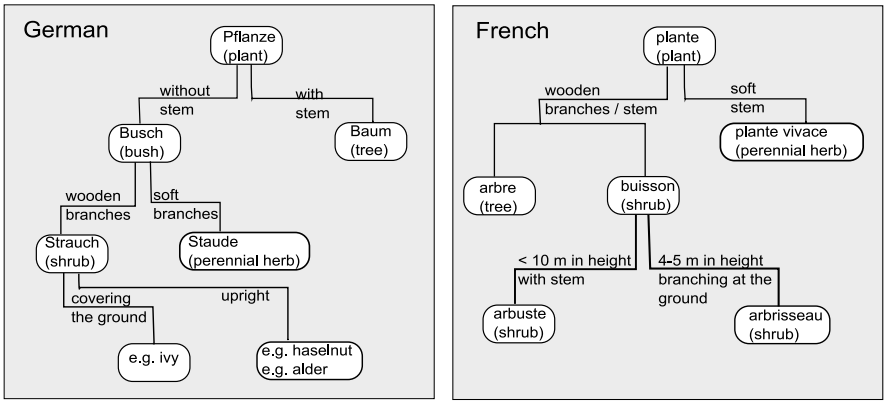

Figure 1. Taxonomic hierarchies in German and French (Source: Meyers Lexikon, LEO Deutsch-Französisch Wörterbuch, http://dict.leo.org/frde) are built based on different classification criteria.

The classification is in many cases ambiguous, depending on the habitat and living conditions of an individual plant. Ivy for instance can grow either covering the ground or climbing at trees.

4. Design of a Bilingual Eco-ontology

The challenges discussed above lead to the following design goals for a bilingual eco-ontology:

- A bilingual eco-ontology will consist of the alignment (also known as mapping) of several ontologies, each consistent in itself. A single ontology with one root is not desirable.
- Mechanisms such as statistical or fuzzy algorithms for handling uncertainties will be introduced. One-to-one translations and alignments are not possible.
- Mechanisms for dynamic updates, maybe even self-learning or self-organization will be provided. A static ontology will soon be outdated (this design goal is not addressed here).

There are different approaches to implement ontologies, among them: Topic Maps (ISO 13250) and ontologies in OWL. While Topic Maps better support "human semantics", OWL (especially OWL DL) supports machine reasoning and automatic search engines, which is a central feature of our project.

While striving for the first two design goals we revert to the way how semantic heterogeneity is usually handled in description logics ontologies (e.g., OWL DL), that is by vocabulary bridges (Catarci and Lenzerini 1993), and extend this mechanism with algorithms for the construction of fuzzy sets which are used to map concepts for which no bridges exist. Different from state-of-the-art bridging mechanisms we consider only bridges which are based on the *equality* of source and target concepts.

A comprehensive summary of recent fuzzy approaches to web applications is given by Nikravesh (2006). Uncertainty and semantic overlap in a single ontology can be modeled using fuzzy degrees of subsumption (Holi and Hyvönen 2006), with a fuzzy extension to description logic (Straccia 1998), or even with fuzzified OWL (Stoilos et al. 2005). Bayesian networks use a sound mathematical (statistical) background for reasoning in fuzzy ontologies. They, however, rely heavily on the availability of probabilistic information from domain experts, which is often not available.

Unlike Bayesian networks we use *algorithms* for dealing with fuzziness. Instead of forcing a crisp one-to-one translation of German and French terms, our algorithm semantically expands both the German and French terms. A fuzzy realm of the original search term is created. We consider

this combination of loosely bridged ontologies with fuzzy mappings, as shown in figure 2, as an original contribution to research.

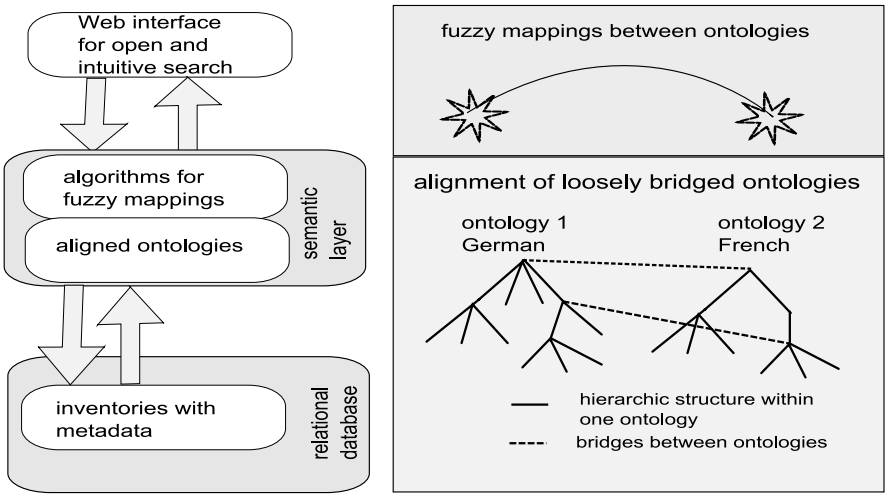

Figure 2. Introducing a semantic layer (consisting of algorithms for fuzzy mappings and aligned ontologies) in an existing eco-database environment.

To sketch the approach in more detail, we assume that the user enters the term "Staude" into the search form and looks for relevant resources in any encoding language (i.e., German or French). Since there is no bridge between the concept denoted by the search term and a semantically equivalent concept in the French ontology, the algorithm climbs the German concept hierarchy until it finds a concept for which a bridge to the French ontology has been defined. In the example this is the concept Pflanze which is defined as being equal to the concept Plante by the bridging axiom Pflanze ≡ Plante. While proceeding, the algorithm adds the names of all concepts visited including their children's to the set of search terms, in the example the terms "Staude", "Busch" and "Strauch".

Starting from the target concept of the bridge, the algorithm then retrieves all subsumed concepts and adds their names to the set of search terms, in the example the terms "arbre", "plante vivace", "buisson", "arbuste", and "arbrisseau". In pseudo code the described algorithm can be put in the following way:

```
Declaration
item:         the given search term
targetitem:   the translated term of item
parent:       hypernym of item (generalization)
child:        hyponym of item (specialization)
```

```
query-terms:  list of search terms. At the beginning
              query-terms contains the search term
              item. At the end of the algorithm it
              contains additional, semantically related
              terms.

Main Program
  if bridge-exists(item) then
     /* add the search term itself */
     add-to-query-terms(item)
     /* add the via bridging translated search term */
     add-to-query-terms(targetitem)
  else query-expand(item)
End Program

Subroutine query-expand(item)
  if NOT bridge-exists(item) then
      /*add the specialized terms of item */
      include-children(item)
      /* back-tracking */
      query-expand(parent)
  else  /*bridge exists */
    /* add the children of the translated search term*/
    include-children(targetitem)
End Subroutine
```

During execution the algorithm constructs a complex concept, in our example Staude ⊔ Busch ⊔ Strauch ⊔ Arbre ⊔ Plante_vivace ⊔ Buisson ⊔ Arbuste ⊔ Arbrisseau.

This concept represents to one part a well-defined extension of the original concept Staude, which assumes that the user is also interested in resources which are closely related to those addressed by the search term (cf. the concept of open search as introduced in the first section). To another part it represents a *fuzzy* extension of the original concept with respect to the certainty that also objects in French corresponding to those denoted by the German "Staude" are included in its (set-theoretic) interpretation: The disjunction of the subsumed concepts does not necessarily have the same extension as the subsuming concept. There might be plants which cannot be classified as either "arbre", "plante vivace", "buisson", "arbuste", or "arbrisseau".

5. Conclusion

Natural language is a highly dynamic system, with significant semantic changes over time and semantic differences between social-cultural

groups. Nevertheless it is the elementary postulate of science, that there is an order in nature (Levi-Strauss 1962), which is reflected by human conceptualizations, that can be modeled as ontologies.

In combining the rigid frame of ontologies with the flexible techniques of fuzzy algorithms, a promising approach to a data structure supporting an open and intuitive search in a bilingual environmental database was designed. The implementation will make use of standard Semantic Web technology offering the possibility of future extensions to the semantic layer in a similar way as described (e.g., inclusion of zoological taxonomies). Since the algorithms are a key component of the architecture, it will be interesting to analyze how modifications thereof will affect recall and precision of predefined queries as well as system performance. Further research on the integration of the eco-ontology into the virtual data base project (Frehner and Brändli 2006) is planned for the near future.

Acknowledgements

The authors sincerely thank Jürg Schenker, Martin Hägeli and Martin Brändli for the fruitful discussions and leadership that made this research possible. This research was funded and conducted in cooperation with the Swiss Federal Office for the Environment (FOEN).

References

Agarwal P, (2005) Ontological considerations. GIScience, International Journal of Geographical Information Science, vol 19, no 5, pp 501-536

Baltensweiler A, Brändli M (2004) Web-based Exploration of Environmental Data and Corresponding Metadata, in Particular Lineage Information. In: Scharl A (ed) Environmental Online Communication. Springer-Verlag, London, pp 127-132

Catarci T, Lenzerini M (1993) Representing and Using Interschema Knowledge in Cooperative Information Systems. Int. J. Cooperative Inf. Syst., 2(4), pp 375–398

Fonseca F, Martin J, Rodríguez MA (2002) From Geo- to Eco-ontologies. In: Geographic Information Science, Second International Conference, GIscience 2002, USA, Egenhofer MJ, Mark DM (eds), Springer LNCS 2478, pp 93-107

Frehner M, Brändli M (2006) Virtual database: Spatial analysis in a Web-based data management system for distributed ecological data. Environmental Modelling & Software, vol 21, pp 1544-1554

Grimm J, Grimm W (1854) Deutsches Wörterbuch von Jacob Grimm und Wilhelm Grimm. 16 Bde. in 32 Teilbänden, Leipzig, Hirzel S (ed) 1854-1960

Gruber TR (1993) A translation Approach to Portable Ontology Specifications. Knowledge Acquisition, vol 5, 199-220

Holi M, Hyvönen E (2006) Modeling Uncertainty in Semantic Web Taxonomies. In: Soft Computing in Ontologies and Semantic Web, Studies in Fuzzyness and Soft Computing, vol 204, Zongmin Ma (ed), ISSN 1434-992, Springer, pp 31-46

Levi-Strauss C (1962) Das Wilde Denken, Shurkamp Taschenbuch Wissenschaft, ISBN 3-518-27614-X, pp334

Nikravesh M (2006) Beyond the Semantic Web: Fuzzy Logic-Based Web Intelligence. In: Soft Computing in Ontologies and Semantic Web, Studies in Fuzzyness and Soft Computing, vol 204, Zongmin Ma (ed), ISSN 1434-992, Springer, pp 149 – 209

Salton G (1970) Automatic Processing of Foreign Language Documents. Journal of the American Society for Information Science, vol 21, no 3, 187-194

Smith B, Varzi A (1999) The Formal Structure of Ecological Contexts. Modeling and Using Context, Second International and Interdisciplinary Conference, 1999

Stoilos G, Stamou G, Tzouvaras V, Pan JZ, Horrocks I (2005) Fuzzy OWL - Uncertainty and the Semantic Web. In: Proceedings of the International workshop on OWL, Experience and Directions (OWL-ED 2005)

Straccia U (1998) A fuzzy description logic. In: Proceedings of AAAI-98, 15th National Conference on Artificial Intelligence, Madison, Wisconsin

Vossen P (2002) EuroWordNet General Document, Version 3, University of Amsterdam, pp108