EnviroInfo 2008 (Lüneburg)
Environmental Informatics and Industrial Ecology
 ISBN: 978-3-8322-7313-2

Improving an Environmental Ontology by Incorporating User-Input

Bettina Bauer-Messmer[1], Thomas Scharrenbach[1] and Rolf Grütter[1]

Abstract

Ontologies supporting open and intuitive search in heterogeneous databases have been studied by various research groups for several years. A crucial factor for success in the use of ontologies is the adjustment of the conceptualization represented in the data to the conceptualizations of the individual users. The generation of a high-quality ontology is very cost intensive. The ontological opening up of the data content is straightforward, whereas the conceptualizations of the users are rather difficult to grasp. We present a novel approach for reinforcement ontology learning integrating user-input. The knowledge base consists of expert ontologies modeled beforehand by experts and user-input will be used to create an user ontology. This paper presents the concept of how to incorporate user-input into existing ontologies.

1. Introduction

1.1 Ontologies in environmental sciences

Processing natural language input from users requires information systems, which "understand" the input up to a certain degree. 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 & Varzi 1999] or they stress the aspect of semantic heterogeneity of source communities [Fonseca et al. 2002]. In this paper the term eco-ontology refers to both aspects, in the sense that the heterogeneous data sources are combined with data about ecological niches. Eco-ontologies represent the environmental domain and parts of the geospatial domain. For the geospatial domain no consensual ontology exists so far, and, as long as the ambiguities within concepts are not clarified, it will, according to [Agarwal 2005], not be possible to create one. Therefore, we use a system of loosely coupled ontologies representing the different conceptualizations of the heterogeneous eco-community (e.g. biologists, geographers, politicians, etc.). The loosely coupled ontologies are linked together and are implemented within a so called semantic layer.

1.2 Open and intuitive search

Public users do not know the internal structure of the database and/or GIS system which implies that systems expecting users to bring thorough internal knowledge are doomed to fail. As a consequence, we construct a system allowing queries without time-consuming user-trainings or manuals, which are never read.

Public users prefer entering their search queries using natural language over learning a machine oriented artificial language. Nevertheless, users tend to adjust their queries to the system up to a certain de-

[1] {bettina.bauer, thomas.scharrenbach, rolf.gruetter}@wsl.ch, Swiss Federal Institute for Forest, Snow and Landscape Research, Zürcherstrasse 111, CH-8903 Birmensdorf, Switzerland

gree. Novice users and children typically enter full sentences in the search bar of search engines, whereas experienced users stick to short and meaningful keywords. It is observed that the use of SMS also influences the use of language. Generally, short synonyms are preferred and furthermore, bilingual people even tend to use these shortcuts in either language. The above mentioned tendencies and user preferences are picked up in our approach to allow natural language user-input, but restricted to collections of search terms in-stead of full sentences. Major attention is given to short forms and synonyms in order to compensate for the users tendency to prefer least "type-intensive" terms.

In the following section, the structure of the DNL data center is presented, in particular the semantic layer. Section 3 presents the method on how to incorporate user-input into the ontology. Section 4 sketches the concept of the implementation. Discussions and outlook conclude this paper.

2. Heterogeneity in environmental data overcome with a semantic layer

2.1 The data center nature and landscape (DNL)

Data of protected biotopes of national relevance in Switzerland is collected in a national database, as required by the Swiss nature and cultural heritage protection act (Natur- und Heimatschutzgesetz NHG). The data collection is organized in inventories with all data about specific types of biotopes like bogs, high moors, dry grasslands, amphibian spawning grounds, etc. Inventories comprise all kinds of data: GIS geometries (polygons or points), attribute data, aerial photographs, object descriptions, control of success, legal documents, political consultations, correspondence, and last but not least metadata. Data collections span several decades, and obviously, the sampling techniques have changed considerably over such long time periods. This makes high quality metadata essential for the correct interpretation and future reuse of existing data.

2.2 The virtual data center (DNL/VDC)

The DNL datacenter is one of the Swiss national environmental databases. Via a mechanism called "virtual datacenter", the DNL is connected to external data sources such as databases for fungi, lichen, faunistic and floral information. The more heterogeneous and complex the "virtual datacenter" becomes, the more important becomes the description of data. Therefore, the datasets are described following the GM03 metadata standard [GM03]. For each DNL inventory a complete set of GM03 metadata exists, which will be also accessible using ontological search facilities [Bischof & Bauer-Messmer 2008]. A comprehensive description of the virtual data center can be found in [Frehner & Brändli 2006].

2.3 Semantic layer in the DNL/VDC

The DNL database was retroactively supplemented with a semantic layer. This semantic layer consists of an application-specific, bilingual eco-ontology and of an algorithm for the semantic expansion of search terms. It also consists of an algorithm for the transformation of search term lists into SQL queries for data access. The idea is to extend the result sets obtained from database queries in a meaningful way by enriching search terms with synonyms, similar terms, further semantically related terms and their translations. This makes accessing the database more open and intuitive for users. The design of the bilingual eco-ontology is given in [Bauer-Messmer & Grütter 2007] and the structure of the semantic layer is discussed in [Grütter et al. 2008].

The bilingual ontology is composed of two independent ontologies in German and French. They are related to each other by means of terminological axioms in terms of equalities such as Moorlandschaft ≡ pay-sages_marécageux. The bridged ontology holds 1,155 items. Because of the many synonyms and similar terms used to label the items, the ontology holds over 2,500 terms. These terms denote thematic notions such as endangered species of animals and plants. They also denote spatial notions such as inventory objects (biotopes) and administrative regions.

The semantic layer is implemented as a Web service. The Web service encapsulates a knowledge base which is created and queried by the OWL reasoner pellet 1.4 [Pellet]. The reasoner is accessed at the Wonder Web OWL API (http://wonderweb.semanticweb.org/). The ontology that is loaded by the reasoner has been constructed using version 3.2 of the Protégé 2000 ontology editor [Protégé]. The language used for the ontology is OWL-DL. The description logic expressivity of the ontology is that of ALCHO (D) In order to figure out how the users conceptualize the domain, a content analysis was performed based on the transcripts of interviews held with users in an early stage of the project. The identified concepts were used as indicators for the terminologies to import in the ontology.

3. Incorporating user-input into the ontology

3.1 Ontology learning

Cognitive learning theories stress the importance of prior knowledge. Learning is connecting new concepts to the existing cognitive structures [Ausubel 1978]. We will apply this constructivistic approach from learning psychology to our ontology. New concepts (i.e. user-input, which cannot be matched with existing concepts in the ontology) must be connected to existing concepts. This process of connecting new concepts to existing semantic structures is highly complex, requires a thorough understanding of both the new concept and the existing concepts and therefore cannot be done automatically.

Following the Web2.0 philosophy knowledge is generated where it is located. The users, who entered the search term which led to a new concept, will be asked to connect it to existing structures, because they know best in which semantic context he or she expects to find it.

3.2 Continuous user specific adjustments of the ontology

The ontology in use was manually generated with collections of typical domain-specific terms from the DNL database. Both, the conceptualizations of the potential users and the conceptualizations represented in the data, were considered during ontology generation. The conceptualization of the expert users is well represented, whereas the conceptualizations of the future public users are quite difficult to grasp. One approach to identify the mental maps and conceptualizations of future users is to let them describe use cases. Use cases, first introduced by Jacobson [Jacobson 1992], define the functional requirements in software engineering and therefore implicitly give information about the users' conceptualizations. The use case approach, however, can only be applied for a limited number of users, since it relies upon the assumption that the user community shares common conceptualizations. This is a reasonable assumption as long as the users share the same cultural and educational background. But already by involving experts from different science domains, the conceptualizations start to diverge. The conceptualization of public users accessing the data via the internet is fairly unknown (Fig. 1).

Not only conceptualizations but also completeness of the ontology is an important issue. Ontologies represent a certain portion of all existing concepts and hence a portion of the vocabulary of a language. It is technically not very challenging to find out, whether a given ontology covers the concepts represented

in a database. However, defining the completeness of an ontology with respect to the conceptualizations of all potential users is quite impossible. One way to overcome this is continuous, user-specific adjustment of the ontology, approximating the completeness of an ontology in the way that users implicitly and explicitly add information to the knowledge base.

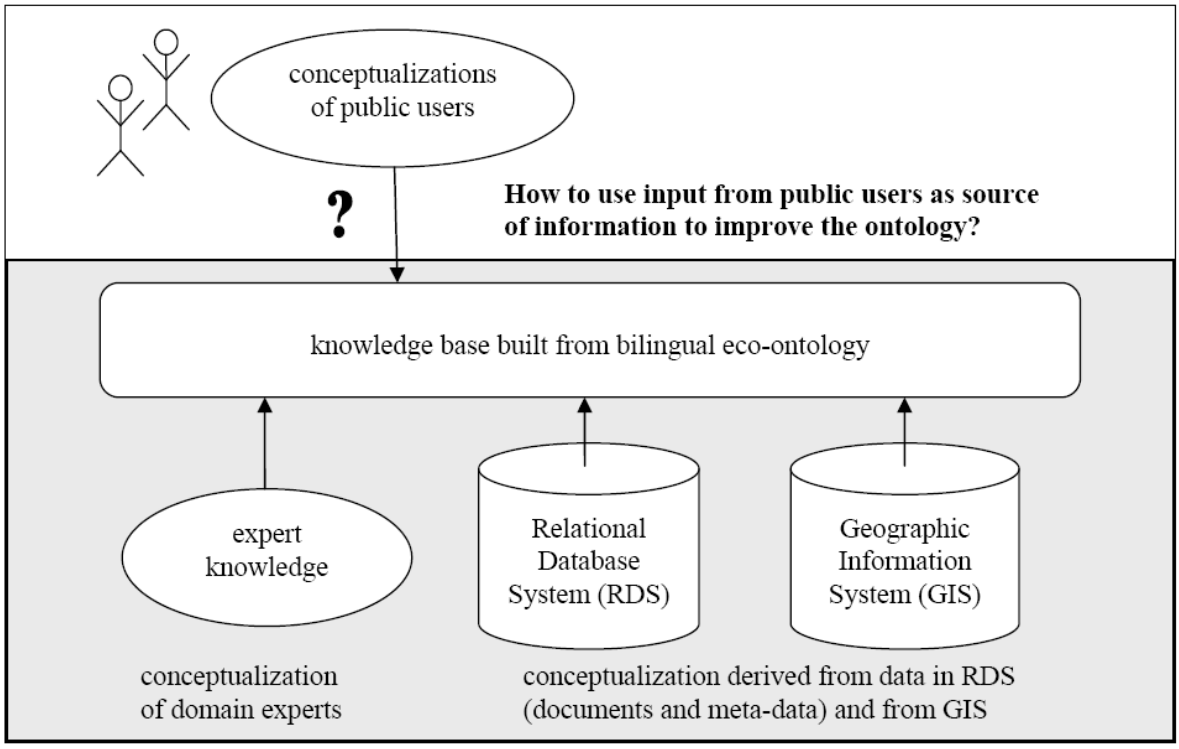

Fig. 1: Conceptualizations of the experts and conceptualizations represented in the data (documents, meta-data and GIS data) build the basis for the manually constructed bilingual eco-ontology used in the semantic layer (up-arrows). The conceptualizations of the public users however are not incorporated yet in the semantic layer (down arrow). This is subject to current research.

3.3 Implicit and explicit user-input

Adjusting ontologies to user conceptualizations is very cost intensive, because the work is typically done manually and requires time consuming user interviews and polls. Users, however, spend a lot of time working with the system, entering search terms and implicitly judging the results by either following the suggested links (or opening the presented documents) or by ignoring the systems suggestions. Why not use this input as a valuable source of information about the user's preferences and dislikes? There are two ways how to incorporate user-input:

1. Explicit user-input is obtained by asking users to place a new concept, derived from a failed search term, into the existing taxonomy. Alternatively explicit user-input is gained by asking about the usefulness of a given response, e.g. the user searches the string "Flachmor" and the system asks "did you mean: Flachmoor?" By clicking yes or ignoring the suggestion the user explicitly evaluates the system.

2. Implicit user feedback can also be used to complete the ontology. When a user enters three search terms and two of the terms are represented in the ontology but the third one is not, then the system can assume a common semantic context of the three terms, and therefore suggests in future queries the third term as potentially relevant term.

A user is not a user. The question of the quality of the user-input remains open. It is assumed that what we call a "Wikipedia-effect" will take place. That is, extraordinary conceptualizations will be cancelled out by "typical" or "average" user-input. A self-regulating effect is expected as observed in other social software environments. All the same, this does not discharge the operators of an ontology based user-service from regularly checking and monitoring the development of the ontology.

4. Concept of implementation

4.1 Combination of expert and user ontology

The knowledge base consists of two parts, the expert ontologies manually constructed by experts and the user ontologies created from user-input. For the expert ontologies, classical logical reasoning will be possible. The combination of expert and user ontologies, however, requires more sophisticated mechanisms to keep the combined ontologies logically consistent. A query to the knowledge base provides in the best case a list of results. In case of too numerous results, only the n best results (n-best list) will be presented to the user. If a search term is not found in the existing ontology the user will be asked to make a suggestion on how to incorporate the term in the existing taxonomy as a new concept. The quality of user-input, especially in internet applications with public users, cannot be guaranteed. As a consequence, it is important to keep the expert ontology separated from the knowledge generated by public users. Only from time to time an update is conducted, where experts first manually check the quality of the user-input and then incorporate those parts of the user-input ontology, which meet their quality requirements into their expert ontology.

4.2 System architecture for user-driven ontology optimization

The virtual data center of the DNL project is implemented as a service oriented architecture (SOA). The semantic layer (i.e. the ontology and its algorithms) is encapsulated in web services, which semantically enhance the user's search terms.

Implicit user-input will be generated by what we call context groups. We assume that the combination of given search terms has a specific meaning in the user's conceptualization and hence form a semantically related group of terms or a context. The incorporation of implicit user-input will be approached using statistical methods.

Explicit user-input is asked whenever a search term is entered that cannot be found in the existing ontology or loosely coupled ontologies. The unknown term is presented to the user with the visualization and editing tool WOW (WSL Ontology Webeditor), which allows the user to browse the ontology. The user is then explicitly asked to add his search term to the ontology by generating links to existing concepts. The WOW editor was developed at WSL in order to allow public users to enter new concepts and connect it to the existing ontology. Technically, the expert part of the ontology is not changed, but rather an additional part is constructed from user-input. Thus, the public user-input is clearly separated from the expert knowledge (Fig. 2). A moderator decides in regular intervals about the integration of public input into the expert ontology.

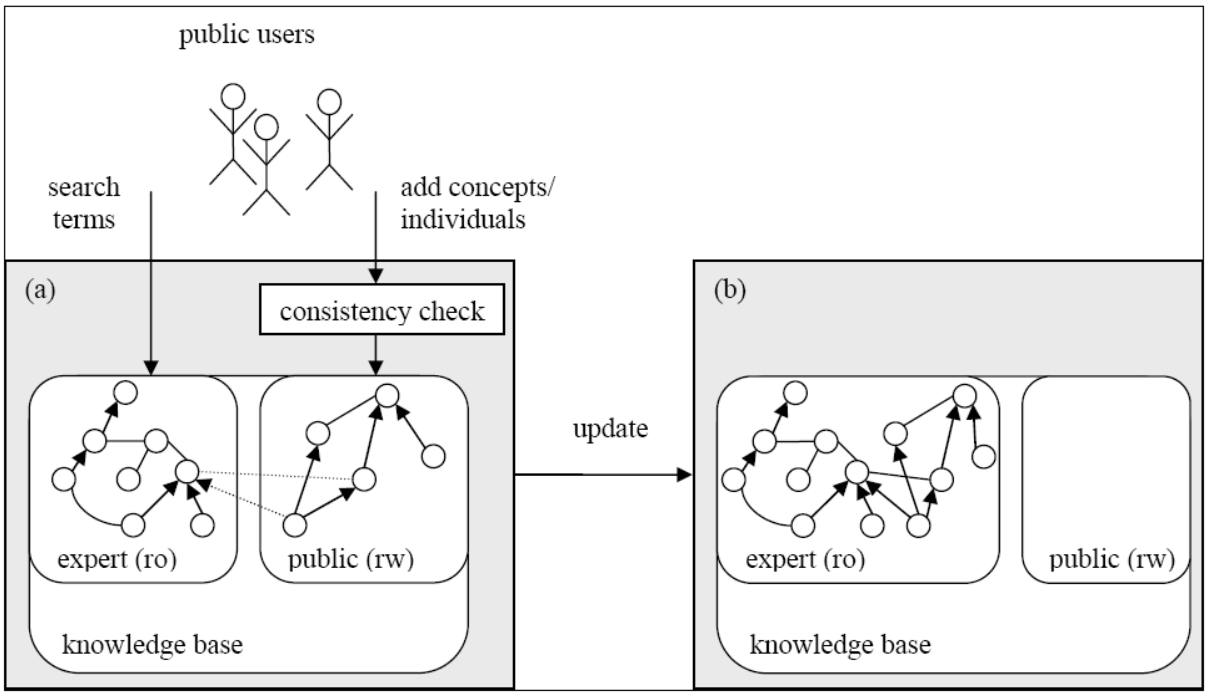

Fig. 2: System architecture based on an expert ontology (bilingual eco-ontology, ro: read only) with a public part (rw: read write), which can be modified by public users (a). In case of a failed search, public users can add new concepts and/or individuals to the ontology using the WOW visualizing and editing tool. New concepts/individuals are only added, if the consistency of the knowledgebase is not endangered. At given intervals experts per-form an update, i.e. they manually check the linguistic consistency of the public ontology. Those parts created from user-input, which fulfill the quality requirements, become part of the expert ontology. After the update (b) the expert ontology has grown and the public part is empty.

Allowing public users to influence such a central component of the system as the ontology always bears risks. The highest risk is expected from users flooding the ontology with trash-data by entering semantically uncorrelated, non-domain-specific search terms. Therefore, monitoring mechanisms and version control, which allow reloading previous ontology-versions, is essential.

4.3 Resolving inconsistencies in the knowledge base

It is expected that different groups of users will tend to develop the ontology in different directions (e.g. public-non-expert users versus students and researchers in environmental sciences, or maybe even biologist versus geographers or engineers). Furthermore, it cannot be guaranteed that users will not introduce contradictions. There are two kinds of inconsistencies to be expected: First inconsistencies at the logic level, which will be detected by the reasoner (in our case by pellet). Secondly, inconsistencies at the linguistic level, i.e. terms from biological taxonomies mixed up with popular terms. Logic inconsistencies are detected before a new concept or individual entered by a public user is stored in the ontology. The WOW editor does not allow storing items which endanger the consistency of the whole knowledgebase.

Linguistic inconsistencies are generated by user-input, which does not fit into the conceptualizations represented in either the expert ontology or the public ontology. Therefore the user-input needs to be evaluated. The evaluation is twofold: On the one hand experts manually evaluate the ontology at given in-

tervals. On the other hand user interaction with new concepts will be logged and statistical analysis will be performed. Based on these evaluations the new concepts are either integrated into the expert ontology or rejected. Updates of the expert ontology, as described above, will be performed at regular intervals.

4.4 Example

The expert ontology comprises terms about biotopes such as bogs, fens, alluvial planes etc. There are hundreds of terms and synonyms stored in the ontology. However the term “swamp” does not show up. If a public user enters “swamp” the term will not be found in the system. First the user is shown a visual representation of the ontology in the WOW tool, which allows them to browse the existing ontology. The user then gets the opportunity to open the WOW editor and enter the term “swamp” in the public ontology and define relations between “swamp” and existing concept in the ontology. Before the “swamp” concept and its relations are stored in the system, a consistency check is run over the whole knowledgebase. In case of violation of the consistency, the new concept is rejected and the user is asked to redefine the new concept or individual.

As soon as the concept “swamp” is stored, all users will see the “swamp” concept as part of the ontology. Technically, users cannot distinguish between expert parts of the ontology and user parts. However, user interaction with the new concept “swamp” will be logged for further analysis. At given times a human expert evaluates the public ontology and the logging information and decides whether the concept “swamp” will become part of the expert ontology or not.

5. Conclusion and Outlook

The WOW tool, allowing visualizing and editing the ontology, was developed at WSL in order to facilitate the incorporation of explicit user-input over the Web. Incorporating implicit user-input is planned for the future. The availability of user-statistics is crucial for the definition of so called context groups of search terms, which are the basis for the incorporation of implicit user-input. During the test phase of the incorporating explicit user-input, statistical data will be collected, which will be used to further process implicit user-input.

The architecture as described above is under construction. Experiments with large numbers of users are planned for the near future. The mechanisms presented here, especially for ontology learning, can be used for user-specific optimizations, by allowing user-specific ontology additions with individual user profiles. The development of the ontology will be closely monitored for a certain time. If the development appears very diffuse and heterogeneous, different user profiles will be introduced.

Acknowledgements

The authors sincerely thank Martin Hägeli (WSL), Jürg Schenker (FOEN) and Prof. Abraham Bernstein (University of Zurich, Switzerland) for the fruitful discussions. This research was funded and conducted in cooperation with the Swiss Federal Office for the Environment (FOEN).

References

[Agarwal 2005] Agarwal, P.: Ontological considerations in GIScience. International Journal of Geographical Information Science, Vol. 19, No. 5, May 2005, pp. 501-536.

[Ausubel 1978] Ausubel, D., Novak, J., Hanesian, H.: Educational Psychology: A Cognitive View (2nd Ed.). 1978, New York: Holt, Rinehart & Winston.

[Bauer-Messmer & Grütter 2007] Bauer-Messmer, B., Grütter, R.: Designing a bilingual eco-ontology for open and intuitive search. Accepted for publication: ITEE 2007, Third International ICSC Symposium - Information Technology on Environmental Engineering, March 29-30 2007, Oldenburg, Germany.

[Bischof & Bauer-Messmer 2008] Bischof, S., B. Bauer-Messmer: Semantic Enhancement of Environmental Metadata, accepted for publication, EnviroInfo 2008, Lüneburg, Germany.

[Fonseca et al. 2002] Fonseca, F. T., Martin, J., Rodríguez, M. A.: From Geo- to Eco-ontologies, In: Geographic Information Science, Second Int. Conference GIScience, Eds. Egenhofer, M. J., Mark, D. M., LNCS, Vol. 2478, pp. 93-107.

[Frehner & Brändli 2006] Frehner, M., Brändli, M.: Virtual database: spatial analysis in a Web-based data management system for distributed ecological data. In: Environmental Modelling & Software, 21, 2006, pp. 1544-1554.

[GM03] Das Schweizer Metadatenmodell, http://www.geocat.ch/GM03_d.htm, KOGIS, Koordination der Geoinformation, c/o Bundesamt für Landestopografie, CH-3003 Bern.

[Gruber 1993] Gruber, T. R.: A translation Approach to Portable Ontology Specifications. Knowledge Acquisition, 1993, Vol 5, pp. 199-220.

[Grütter et al. 2008] Grütter, R., Bauer-Messmer, B., Frehner, M.: First Experiences with an Ontology-Based Search for Environmental Data. In Proceedings of the 11th AGILE International Conference on GI Science, 6-8 May 2008, Girona, Spain.

[Jacobson 1992] Jacobson, I.: Object-Oriented Software Engineering, Addison Wesley Professional, ISBN 0-201-54435.

[Pellet] http://pellet.owldl.com/

[Protégé] http://protege.stanford.edu/

[Smith & Varzi 1999] Smith, B., Varzi, A.: The Formal Structure of Ecological Contexts. Modeling and Using Context, In Proceedings of the Second International and Interdisciplinary Conference on Modeling and Using Context, Trento, Italy, 9-11 September 1999, LNCS, Vol. 1688 , pp. 339-350.