# First Experiences with an Ontology-Based Search for Environmental Data

Rolf Grütter, Bettina Bauer-Messmer, Marcel Frehner

Swiss Federal Research Institute WSL, An Institute of the ETH Board,
Zürcherstrasse 111, CH-8903 Birmensdorf

**Abstract.** This paper reports on first experiences with an ontology-based search for environmental data in a productive database. We show in a test case how the system supports a user in answering questions that arise in daily business. Because of the complex structure of the process-oriented database the implementation was quite challenging but successful. The use of a domain ontology composed of two component ontologies in German and French proves to be the right design decision in a bilingual context. The presented approach is discussed in relation to existing work. It is concluded that the search space must be extended before evaluating the system by comparing measures like recall and precision with a benchmark.

## 1 INTRODUCTION

The Datacenter Nature and Landscape (DNL) of the Swiss Federal Office for the Environment (FOEN) is operated by the Swiss Federal Institute for Forest, Snow and Landscape Research (WSL). Its core is a relational database system implementing a process-oriented data model. The database holds roughly 200,000 data records which are grouped into 12 inventories of different kinds of biotopes such as bogs, mires, moorlands, alluvial zones, amphibian spawning grounds. The number of objects per inventory ranges from some tens to thousands. Objects and inventories are described in text fields, data tables and in documents stored as binary large objects using a number of specific terminologies (e.g., botanic, zoological). The DNL datacenter contributes to the Virtual Data Center (VDC) which is a system of loosely coupled geospatial databases (Frehner and Brändli, 2006).

Until now, the DNL datacenter has been exclusively accessible to expert users with previous knowledge of the data model and the terminologies used. Following the decision of the FOEN to make parts of it available to the public as part of the future Swiss National Spatial Data Infrastructure, a project aimed at providing an open and intuitive, Web-based access to the DNL datacenter was initiated. The ontology-based search presented in this paper is a first result of this project.

The presented approach to ontology-based query processing has similarities with that of Necib and Freytag (2003). Both attempt to extend the result set obtained from a single relational database in a meaningful way. Necib and Freytag use knowledge encoded in an ontology in order to transform a query in terms of an SQL statement into a semantically equivalent SQL statement enriched with synonyms and more specific terms. Although the transformed SQL statement looks quite similar to that generated during query processing in the herein presented application, Necib and Freytag assume that a user is SQL literate which we do not.

The paper is organized as follows: Section 2 provides an overview of the system architecture. Section 3 explains how the ontology is structured and which steps are taken when processing a query. In the course of a preliminary qualitative evaluation, section 4 describes a first test case which, together with two related approaches, provides input to the discussion in section 5. Section 6 concludes the paper.

## 2 SYSTEM ARCHITECTURE

An overview of the architecture of the VDC is shown in figure 1. The ontology-based search is implemented for the DNL database. The Web services depicted with dashed lines for the Centre Suisse de Cartographie de la Faune (CSCF) and other databases do not exist yet. The grey shaded boxes at the bottom of the sketch show relational databases. The grey shaded box in the middle of the figure represents the semantic layer. The semantic layer consists of a Web service for the semantic expansion of the search terms and of Web services generating database specific SQL queries. The Web service for semantic expansion can work with different and even with multiple ontologies. In contrast, the Web services which generate SQL queries are restricted to a single data source and cannot be reused in other contexts.

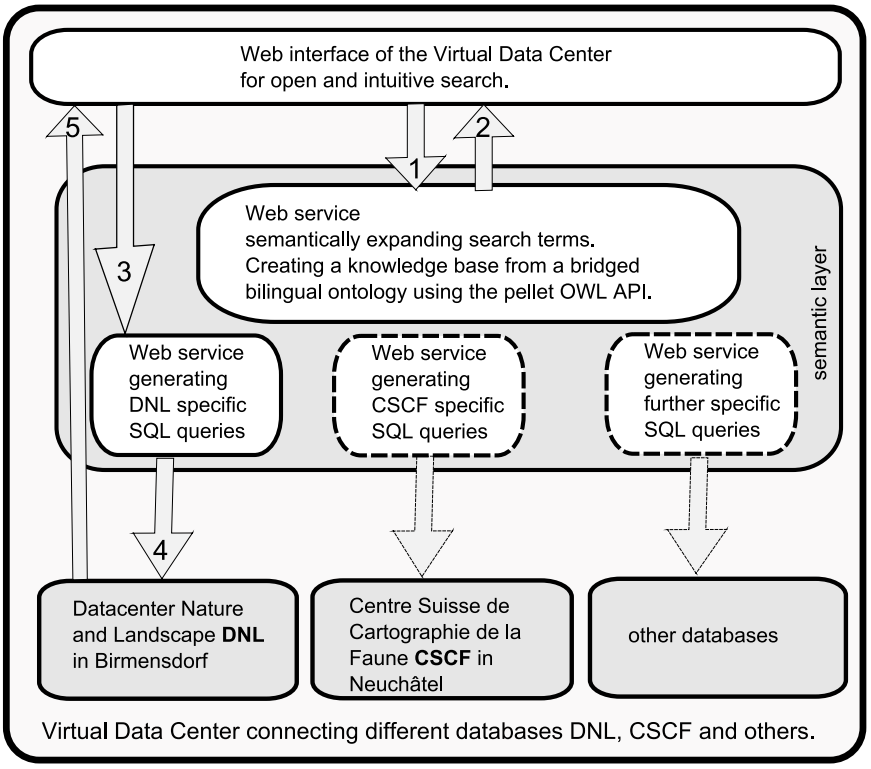


***Figure 1:*** Service oriented architecture of an ontology-based search in the Virtual Data Center

The white box at the top of the figure indicates the user interface of the Web application. The steps that are taken when processing a query are described in section 3.[1]

The Web service that expands the search terms semantically encapsulates a knowledge base which is created and queried by the OWL reasoner pellet 1.4. The reasoner is accessed at the WonderWeb OWL API.[2] The ontology that is loaded by the reasoner has been constructed using version 3.2 of the Protégé 2000 ontology editor. The language used for the ontology is OWL DL. The description logic expressivity of the ontology is that of AΛXHO(Δ). Since the majority of documents and metadata records in the DNL datacenter is written either in German or French the ontology provides a vocabulary in German and French (cf. Bauer-Messmer and Grütter, 2007). In order to figure out how the users conceptualize the domain, we made a content analysis of 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 ONTOLOGY-BASED QUERY PROCESSING

### 3.1 Structure of the Ontology

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 ≡ paysages_marécageux. The bridged ontology holds 1,155 items. These items refer to OWL classes, OWL properties and OWL individuals.[3] The names of these items as well as synonyms and similar terms are represented as values of label properties. Not the syntactically constrained item names are thus the vocables but the label values. These label values are all nouns (i.e. *nomina substantiva*) in nominative singular and nominative plural. We only discriminate between proper names, common names and taxons (cf. table 1).

*Proper names* are used to label individuals. They are normally not translated and the unique name assumption (UNA) holds.[4] An example taken from the test case in section 4 is "Plaun Segnas Sut". As can be seen from this example, names may include several words.

*Common names* are used to label non-taxonomic classes and properties. They are normally translated and the UNA does not hold. Examples taken from the test case in section 4 are "Moorlandschaft" (moorlands) and "Geometrie" (geometry).

*Taxons* are used to label classes in taxonomies. They are normally translated, still the UNA holds: we consider the translated terms as similar terms and not as synonyms. An example is "Castor" which is the latin taxon for the beaver genus.

Since labels do not only store synonyms but also similar terms, a search term matching a label value of an item experiences a moderate semantic expansion with the readout of the label values. For instance, the terms "paysage marécageux" (moorland) and "marais" (mire) with which the item paysage_marécageux in the French ontology is labeled do not share the same extension. Moorlands contain mires but also non-mire areas. Conversely, some mires are located outside a moorland. Note

[1] In the current prototype the semantic expansion of search terms and the generation of SQL queries are implemented as a single service.

[2] http://wonderweb.semanticweb.org/

[3] We will use the terms "class", "property" and "individual" without prefixed "OWL" in the remaining text.

[4] As they are not translated, it is sufficient to assert individuals in a single ontology. We assert them in the German ontology.

that we do not refer to this if we use the term "semantic expansion" in this paper. We rather refer to logical inferences drawn by a reasoner operating on the bridged ontology (cf. next section).

| | Item | Translation | UNA | Example |
|---|---|---|---|---|
| Proper name | individual | no | yes | "Plaun Segnas Sut" |
| Common name | non-taxonomic class, property | yes | no | "Moorlandschaft" |
| Taxon | class in taxonomy | yes | yes | "Castor" |

Table 1: Names and their implementation in the ontology

### 3.2 ONTOLOGY-BASED QUERY PROCESSING

Ontology-based query processing involves a sequence of (pre-) processing steps which are described in this section. We consider the situation where a user enters one or more search terms into the search form and submits the query. The search form provides a field for text entry and two buttons, one for query expansion, another for query evaluation (cf. figure 2).

*1. The input is analyzed.* The input character string is cut into coherent substrings and each of these substrings is compared with the vocabulary of the ontology. The vocables that match any of the substrings are added to a set of terms which is the source data structure for all further processing. The matches are case-insensitive and insensitive w.r.t. number (singular or plural). The number of input terms is not limited.

*2. Each term is semantically expanded.* The expansion depends on the type of the term (class, property or individual) and on the conceptual structure specified by the ontology. Based on the assumption that in most cases the user is looking for individuals we apply the rules: Classes that are not undermost subclasses are expanded to subclasses, undermost subclasses are expanded to individuals, properties are expanded to domain individuals, individuals are expanded to types (i.e. classes). The names of the expanded items together with synonyms and similar terms are rendered as an alphabetically ordered list of checkboxes at the user interface (cf. figure 2).

*3. The terms are logically connected to each other.*[5] Before querying the database the terms in the source data structure and, optionally, the terms selected from the list of expanded items are connected to each other through the logical connectives AND, OR. Depending on the type of the term we apply the rules:[6]

If the term is a class name then it is connected to each of the subclass names through the AND connective and the subclass names are connected to each other through the OR connective. Motivation for the AND connective: The searched individual is a member (the searched individuals are members) of both the class and the subclass.

If the term is the name of an undermost subclass then it is connected to each of the individual names through the AND connective and the individual names are connected to each other through the OR connective. Motivation for the AND connective: The individuals can be understood as nominals collectively constituting an anonymous class. This anonymous class is equivalent to the named

[5] The use of the system will show whether the connections are well defined or need to be modified.

[6] Examples for rules are provided in the next section.

subclass. The searched individual is a member (the searched individuals are members) of both the anonymous class and the subclass.[7]

If the term is a property name then it is connected to each of the individual names through the AND connective and the individual names are connected to each other through the OR connective. Motivation for the AND connective: The individuals can be understood as nominals collectively constituting an anonymous class. This anonymous class is equivalent to the (possibly pseudo) subclass of the domain (which is a named class) defined by the property and the domain. The searched individual is a member (the searched individuals are members) of both the anonymous class and the subclass. Since the property is *by definition* connected to the domain through the AND connective and the anonymous class is equivalent to a subclass of the domain, the property is connected to the individual(s) also through the AND connective.[8]

If the term is an individual name then it is connected to each of the type names through the AND connective. The type names are connected to each other through the OR connective. Motivation for the AND connective: The individual can be understood as a nominal belonging to one or more anonymous classes. These classes are equivalent to the named types. The individual is a member of both the anonymous classes and the types. Two more rules:

- Terms that are entered in the search form are connected to each other through the AND connective.
- Synonyms and similar terms are connected to each other through the OR connective.

The application of these rules has as a result that the WHERE clause of the SQL statement with which the database is queried (cf. step 4) is a formula in conjunctive normal form (CNF). The literals of this formula are the vocables which match the search terms, the synonyms of these vocables, similar terms and – except for the individuals (cf. section 3.1) – their translations. This formula is the intermediate data structure of the process.

*4. The database is queried.* Using the SQL statement generated in step 3 the database is queried. For the prototype we defined a comment field in a data table, a text field in another data table and an extension field in a third data table as a search space, the last in order to distinguish sets with data about objects from data sets with documents. Taken together, these data fields are the target data structure.

According to the categorization introduced by Efthimiadis (1996), the described query expansion is *interactive*, *based on knowledge structures* and *collection dependent*: The user defines the scope of expansion by selecting one or more terms from a list. The class and property hierarchies in the ontology provide a knowledge structure which is dependent on the kinds of data held in the database: The ontology specifies a conceptualization of the domain to which the data in the database belong.

## 4 A FIRST TEST CASE

Together with potential users we compiled a number of use cases before designing the ontology-based search. In one case, the Federal Office for the Environment (FOEN) seeks to answer the question: “Has the geometry of the object named ‘Plaun Segnas Sut’ in the inventory of moorlands changed along with the various revisions (and, if yes, which is the course of change)?” – The question is thus

[7] This is a special case of the first where the subclass is a pseudo subclass.

[8] This is the case of a definition: *Definitio fit per genus proximum et differentiam specificam.*

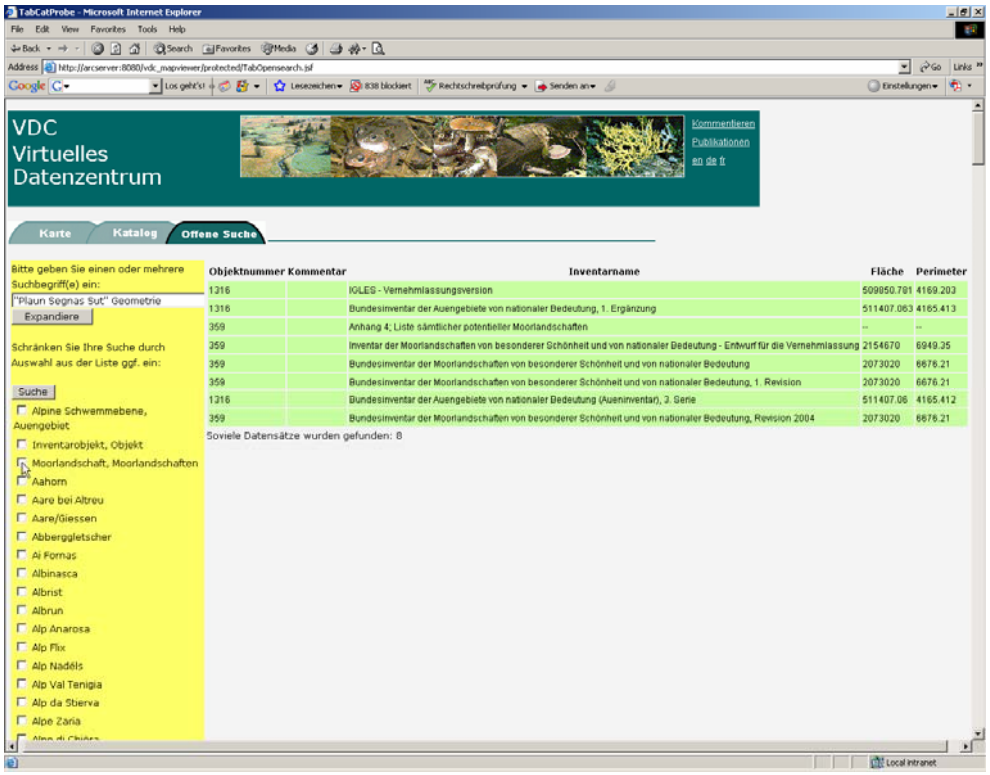


***Figure 2:*** Ontology-based search without semantic expansion (screenshot)

about the geometry of "Plaun Segnas Sut" which is a moorland. Accordingly, we assume a (German speaking) user to enter one of the terms "Plaun Segnas Sut", "Geometrie" (geometry), "Moorlandschaft" (moorlands), either alone or in combination into the search form. Provided the order of entry does not matter, seven entry options can be differentiated: each term alone (three options), two terms in combination (three options), three terms in combination (one option). In order to demonstrate the operation of the ontology-based search we discuss the case where the user enters "Plaun Segnas Sut" and "Geometrie" in the search form. Note that in the remaining cases the user is supported in the processing of the queries in a similar way as demonstrated in the example.

After submission of the query the search terms are semantically expanded (cf. list with checkboxes on the left-hand side of the user interface in figure 2): "Plaun Segnas Sut" expands to "Alpine Schwemmebene, Auengebiet" (alluvial zone), "Inventarobjekt, Objekt" (inventory object) and "Moorlandschaft, Moorlandschaften" (moorlands), which are class names. The reason for this is that an individual named "Plaun Segnas Sut" is a member of either of these classes. "Geometrie" expands to all objects asserted in the ontology ("Aahorn", "Aare bei Altreu" and so forth). Each of them has a geometry in terms of an area and a perimeter.

Searching without selecting any terms from the list of expansions returns eight data sets. As can be seen on the right-hand side of the user interface in figure 2 not all refer to the moorland "Plaun Segnas Sut". Some refer to an object of the same name in the inventory of alluvial zones which is identified by a different object number, namely 1316 instead of 359. The semantic expansion anticipates this result by providing the user with the option to narrow their search. By selecting one or

more terms from the list of expansions the selected terms are added to the query set. Since the use case is about a *moorland* named "Plaun Segnas Sut", the user is supposed to select the terms "Moorlandschaft, Moorlandschaften" from the list. Searching with these terms returns those five data sets with which the question raised in the use case can be answered: Both area and perimeter of the object named "Plaun Segnas Sut" in the inventory of moorlands were cut down between consultation and ordinance but are stable since then.

## 5 DISCUSSION

Lutz and Klien (2006) use a *hybrid ontology approach* for ontology-based retrieval of geographic information. According to the classification introduced by Wache and co-authors (2001) a hybrid ontology approach assumes a shared vocabulary containing the basic terms of a domain which require no formal definitions. These terms are used to build the complex terms of a source ontology for the description of an information source. In contrast, we apply a *single ontology approach* for query expansion. This does not exclude that the "single" domain ontology is composed of (language-specific) component ontologies. Our approach is chosen because the descriptions in the text fields of the databases and the documents have been created with a very similar view on the domain which is also shared by the users. We attempt to request by default all services of the VDC when processing a query and to merge the result sets in a post-processing step. As a consequence service discovery and information retrieval conflate. This makes the architecture of the system much simpler than that presented by Klien and co-authors (2004).

Viegas and Soares (2007) present a hybrid ontology approach using multiple domain ontologies for diverse user communities together with a single application ontology for a geographic database which, however, does not apply a service-oriented architecture. Different from the herein presented approach the authors use the default OWL reasoner provided by the Jena Semantic Web framework which does not provide complete description logic reasoning (Reynolds, 2007). Since we plan to upgrade the reasoning services in the future, complete description logic reasoning is a requirement for our application. Another difference is the way how queries are formulated. After the selection of a user interface, Viegas and Soares provide the user with specific terms for query formulation which are taken from the related domain ontology. This is similar to the *simple query* presented by Klien and co-authors (2004) where the user chooses a concept from an existing application ontology for their query. Viegas and Soares visualize the result sets with a Web-based mapping application. This is also foreseen for spatial data sets in our application.

The bilingual ontology is composed of two independent ontologies in German and French (cf. section 3). This modular design accounts for a very comfortable feature during query processing: The dialog language at the user interface is "automatically" specified by the language in which the search terms are entered. The terms that are listed for selection at the user interface and which semantically expand the search terms are always displayed in the same language as the search terms. This is because the algorithm consults only that ontology for expansion in which the matches are found. Only for the construction of the database query the bridged ontology is consulted. While programming the algorithm we further took care of keeping it independent of the given ontologies. As a result of this approach, further ontologies, for instance one in Italian, can easily be plugged in without changing the algorithm.

As explained in section 3, the WHERE clause of an SQL statement generated in the course of an ontology-based search is a formula in CNF. The literals in this formula are the (possibly expanded) search terms. If we think of the text fields which define the search space as of a junk (or set) of terms, we can say that the formula is satisfied in the text fields of the returned data sets. It is not satisfied in the text fields of the data sets that are not returned. Evaluating a query thus generates two discrete kinds of data sets (one of which might be empty). There is nothing in-between such as a degree of satisfaction. As a consequence, the answer set is not ranked according to relevance as it is the case with today's syntactical search engines but sorted according to a predefined criterion. We use the

process number which is a sequential number and reflects the chronology of data acquisition. Thus the most recent entries (which are not always the most relevant) are displayed on top of the list. Relevance ranking requires a post-processing of the answer set which is outside the scope of the current project but is a possible future extension.

With increasing size the knowledge base created from the ontology becomes more and more able to answer certain questions without querying the database. Preprocessing the term "Moorlandschaft", for instance, lists all individuals which are inventory objects and moorlands. This is an answer to the question "Which moorlands are covered by the National Inventory of Moorlands?" With the current size of the ontology this process is much faster than database queries where a number of data tables must be joined in order to evaluate the query. It will be interesting to observe up to which size the knowledge base outperforms the database. In either case, we do not attempt to replicate the database with the knowledge base. The representation in the ontology will be limited to vocabulary. Data values such as the area and perimeter of an object will not be represented. As a consequence, there will always be questions – such as the one posed in section 4 – which can only by answered by querying the database.

## 6 CONCLUSION

This paper shows in a test case how an ontology-based search supports a user in answering questions that arise in daily business. Because of the complex structure of the process-oriented database the implementation was quite challenging but successful. The use of a domain ontology composed of two component ontologies in German and French proves to be the right design decision in a bilingual context.

In the current application, the ontology plays a twofold role in the processing of queries. It is used for analyzing the search entry and for expanding the search terms semantically. In the first role the ontology acts as a filter: Search terms that are not represented in the ontology are excluded from search. This has some implications for the further development of the application:

1. There must be a possibility to bypass the ontology in case the user enters a term which is not represented but still matches data sets when querying the database.
2. The ontology must be as comprehensive as possible in order to provide with a good chance the terms with which users conceptualize the domain to which the data of DNL belong.
3. Mismatching terms must be captured and frequent mismatches incorporated in the ontology as it evolves.

The first implication is not yet anticipated by the current application. However, user access to the Web service generating DNL specific SQL queries (cf. figure 1) without requesting the semantic expansion service should be easy to implement. The second implication urges us to update the ontology primarily with zoological and botanic terminologies which are only roughly represented in the current version. Along with the extension of the ontology also the search space must be extended by making the documents in the databases accessible to text processing. This will allow for a more systematic evaluation of the semantic search by comparing recall and precision with those of a syntactic search. The third implication draws a bow to a recent project about self-learning and self-organizing ontologies as well as their interconnections which is performed at the Swiss Federal Research Institute WSL.

### Acknowledgements

The authors sincerely thank Jürg Schenker and Martin Hägeli for the fruitful discussions and leadership that made this research possible. They also acknowledge the qualified assistance of Angéline Bedolla in editing the French ontology. This research has been funded and conducted in

cooperation with the Swiss Federal Office for the Environment (FOEN). Related research was funded by the European Commission and by the Swiss Federal Office for Education and Science within the 6th-Framework Programme project REWERSE number-506779 (cf. http://rewerse.net).

## BIBLIOGRAPHY

Bauer-Messmer, B., Grütter, R., Designing a Bilingual Eco-ontology for Open and Intuitive Search. In J. M. Gómez, M. Sonnenschein, M. Müller, H. Welsch, C. Rautenstrauch (eds.). Information Technologies in Environmental Engineering. Springer Verlag: 143–152, 2007.

Efthimiadis, E.N., Query Expansion. In M.E. Williams (ed.). Annual Review of Information Systems and Technology (ARIST), 31, 121–187, 1996.

Frehner, M., Brändli, M., Virtual Database: Spatial Analysis in a Web-based Data Management System for Distributed Ecological Data. Environmental Modelling & Software, 21, 1544–1554, 2006.

Klien, E., Einspanier, U., Lutz, M., Hübner, S., An Architecture for Ontology-Based Discovery and Retrieval of Geographic Information. In F. Toppen, M. Painho (eds.). Proceedings of the 7th Conference on Geographic Information Science (AGILE 2004), Heraklion, Greece, pp. 179–188.

Lutz, M., Klien, E., Ontology-Based Retrieval of Geographic Information. International Journal of Geographical Information Science 20(3), 233–260, 2006.

Necib, C.B., Freytag J.-C., Ontology based Query Processing in Database Management Systems. In Proceedings of the International Conference on Ontologies, Databases and Applications of SEmantics (ODBASE'03), Catania, Italy, 2003.

Reynolds, D., 2007 (December 7) Jena 2 Inference support: The OWL reasoner. http://jena.sourceforge.net/inference/index.html

Viegas, R., Soares, V., Querying a Geographic Database using an Ontology-Based Methodology. In C.A. Davis, A.M. Vieira Monteiro (eds.). Advances in Geoinformatics: VIII Brazilian Symposium on Geoinformatics, GeoInfo 2006. Springer-Verlag: 121-136, 2007.

Wache, H., Vögele, T., Visser, U., Stuckenschmidt, H., Schuster, G., Neumann, H., Hübner, S., 2001 Ontology-based integration of information—a survey of existing approaches. In IJCAI-01 Workshop: Ontologies and Information Sharing, Seattle, WA, pp. 108–117.