An Architecture for Ontology-Based Discovery and Retrieval of Geographic Information

Eva Klien[1], Udo Einspanier[1], Michael Lutz[1] and Sebastian Hübner[2]
[1]Institute for Geoinformatics (IfGI)
University of Münster, Germany
{klien|spanier|m.lutz}@uni-muenster.de

[2]Intelligent Systems Group
University of Bremen, Germany
huebner@informatik.uni-bremen.de

SUMMARY

Finding and accessing suitable information in the open and distributed environments of current Spatial Data Infrastructures (SDIs) is a crucial task. Catalogues provide searchable repositories of information descriptions, but the mechanisms to support the tasks of discovery and retrieval are still insufficient. Problems of semantic heterogeneity caused by synonyms and homonyms can arise during free-text search in catalogues. Moreover, once a suitable Web Feature Service (WFS) is found and accessed, the property names of a feature are often difficult to interpret. This paper introduces an architecture for ontology-based discovery and retrieval of geographic information that solves semantic heterogeneity problems of current query capabilities. Based on a (real-world) scenario from the area of flood management, the application of our approach shows that the information requestor can be efficiently supported.

KEYWORDS: *semantic heterogeneity, ontologies, GI discovery, GI retrieval*

INTRODUCTION

Geographic information (GI) is the key to effective planning and decision-making in a variety of application domains. So-called intelligent web services permit easy access and effective exploitation of distributed geographic information for all citizens, professionals, and decision-makers (Bishr, 2000; Brox, 2002).

This paper focuses on the discovery and retrieval of geographic information. The specifications provided by the OpenGIS-Consortium (OGC) enable syntactic interoperability and cataloguing of geographic information. However, while OGC-compliant catalogues support discovery, organisation, and access of geographic information, they do not yet provide methods for overcoming problems of semantic heterogeneity. These problems still present challenges for the GI discovery and retrieval in the open and distributed environments of Spatial Data Infrastructures (SDIs).

One possible approach to overcome the problem of semantic heterogeneity is the explication of knowledge by means of ontologies, which can be used for the identification and association of semantically corresponding concepts (Wache, 2001). In this paper we introduce an architecture for ontology-based discovery and retrieval of geographic information. To this end, we extend the query capabilities currently offered by OGC-compliant catalogues with terminological reasoning on metadata provided by an ontology-based reasoning component. We show how this approach can contribute to solve semantic heterogeneity problems during free-text search in catalogues and how it can support intuitive information access once an appropriate resource has been found.

MOTIVATING EXAMPLE: DISCOVERING INFORMATION ON WATER LEVELS IN THE ELBE RIVER

Throughout this paper we use a motivating example to illustrate semantic heterogeneity problems that can occur when using state-of-the-art GI query possibilities and our approach to their solution. However, our work is designed to be independent of a particular GI domain and is not restricted to only this example.

John is a hydrologist who is interested in water levels of the Elbe River. As an expert in the field he knows the existing control points in the river. He wants to know the measurement of the water level at a specific control point at a specified time. Since John does not know about an existing Web Feature Service (WFS) offering this kind of information, he makes use of an OGC-compliant catalogue in order to find appropriate information for answering his question: "*What is the water level at control point X at time Y in the Elbe River?*"

There are several data providers that offer information about water levels in the Elbe River via a standardized WFS interface[1].

a) The Federal Agency for Hydrology (Bundesanstalt für Gewässerkunde, **BafG**):
b) The Electronical Information System for Waterways (Elektronisches Wasserstraßen-Informationssystem, **ELWIS**):
c) The Czech Hydrometeorological Institute (**CHMI**):

Table 1 lists the names of the GML features returned by these WFS and their property names.

Table 1: Names of the GML features returned by three WFS and their properties

WFS	BafG	ELWIS	CHMI
Feature	Pegelmessung	WasserstandMessung	StavVody
Name of the control point	name	pegel	stanice
Unique ID of the control point		id	
Internet address of the control point	url	quelle	url
Water level measured in cm	wasserstand_cm	hoehe	stav
Date and time of the measurement	zeitpunkt		datum
Date of the measurement		datum	
Time of the measurement		uhrzeit	
Geometry as Point	gml:pointProperty	standort	gml:position
Name of the river			tok
Discharge in cubic meters per second			prutok

[1] These agencies do not yet provide their data on water levels through WFS but through normal html pages. For implementing the example the information is parsed from these pages and provided through WFS interfaces. The access points to these services are provided at http://www.meanings.de/

SEMANTIC HETEROGENEITY PROBLEMS DURING STATE-OF-THE-ART DISCOVERY AND RETRIEVAL OF GEOGRAPHIC INFORMATION

This section describes possible problems caused by semantic heterogeneity between user requests and application schemata or metadata descriptions, if users and providers of geographic information are from different information communities (OGC, 1999).

Discovery

In current standards-based catalogues (e.g. GDI-NRW (2002)) users can formulate queries using keywords and/or spatial filters. The metadata fields that can be included in the query depend on the metadata schema used (e.g. ISO 19115) and on the query functionality of the service that is used for accessing the metadata.

Two types of semantic heterogeneity can lead to problems if John performs a simple keyword-based search, e.g. using the terms "water level" and/or "Elbe". These types are classified by Bishr (1998) as:

1. *Naming heterogeneity (synonyms):* John may fail to find the existing WFS that are offering the information, because their metadata description contains slightly different terminology, e.g. "depth" or "Labe" (the Czech name for "Elbe").

2. *Cognitive heterogeneity (homonyms)*: John's request could also result in finding services that are not appropriate for answering his question, thus indicating the occurrence of *cognitive heterogeneity*. This would be the case, e.g., if John's free-text search for "water level" resulted in discovering a service for depicting the network of water level control points, without the actual information about the current water level or a service providing *groundwater* rather than *surface water* levels.

These examples show that keywords used in free-text entries have to be considered a poor way to capture the semantics of a query or item (Bernstein, 2002).

Retrieval

Another major difficulty arises if John wants to access geographic information via one of the existing WFS. The `DescribeFeatureType` request (OGC, 2002) returns the application schema for the feature type, which is essential for formulating a filter for the query. John now runs into troubles if the property names are not intuitively interpretable (cf. Table 1). For example, he can only guess that the property names "hoehe" (German) or "stav" (Czech) refer to the measurement of the water level. The goal of the architecture presented in this paper is to provide user support for interpreting property names adequately, since this is a precondition for formulating an appropriate query.

ONTOLOGY-BASED APPROACH

To solve the semantic heterogeneity and interpretation problems presented in the previous section, an approach is needed that exceeds the capabilities of current free-text search facilities in catalogues and supports an intuitive interpretation of property names. Accepting the diversity of geographic application domains, such an approach would need to enable navigating differences in meaning (Harvey, 1999). Stuckenschmidt (2002) suggests to use explicit context models that can be used to re-interpret information in the context of a new application. Ontologies have become popular in information science as they can be used to explicate contextual information.

We adopt a modified version of Gruber's (Gruber, 1993) often-quoted definition of the term "ontology" by Studer (1998), who defines it as "an explicit formal specification of a shared conceptualization" (a conceptualization being a way of thinking about some domain (Uschold,

1998)). This makes the ontology a perfect candidate for communicating a shared and common understanding of some domain across people and computers (Studer, 1998).

Hybrid Ontology Approach

The hybrid ontology approach used in our architecture for enhancing information discovery and retrieval has been adopted from Visser (2002). It is based on the idea of having a source-independent shared vocabulary for each domain (Figure 1).

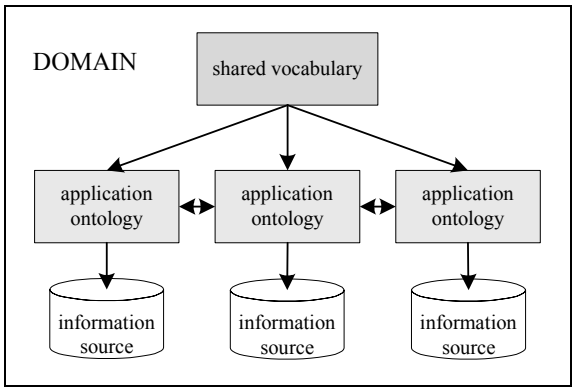

Figure 1: The hybrid ontology approach (Visser, 2002), modified.

It is assumed that the members of a domain share a common understanding of certain concepts. These concepts require no further explication and therefore form the basic terms contained in a *shared vocabulary*. Once a shared vocabulary exists, the terms can be used to make the contextual information of an information sources explicit, i.e. to build an application ontology for it (Visser, 2002). Thus, the vocabulary has to be general enough to be used across all information sources that are to be annotated within the domain, but specific enough to make meaningful definitions possible (Schuster, 2001). The task of constructing an application ontology lies in the responsibility of the provider of the information source.

For the ontology-based annotation of information sources we have made two modifications to this approach. First, the information sources are not annotated directly. Instead, we describe the feature type provided by a service, which is defined through its application schema. To be more precise, the shared vocabulary is used to describe in detail the properties included in the schema. There is therefore an additional level of semantic annotation. Together with the syntax, which can be requested via the normal `DescribeFeatureType()` operation, the annotation of an information source is complete. Second, we do not only use domain-specific ontologies (e.g. measurements, hydrology), but also domain-independent ontologies (e.g. SI units).

In our example, John searches for the water level of control point X at time Y. In our approach, this means that he uses these properties (i.e. “location”, “water level”, and “date and time of measurement”) to describe a feature type. The precise formulation of the query and its execution and result are presented in the next section.

Ontology-Based Search

We distinguish two (closely-related) types of query. In a *simple query* the user can choose a concept from an existing application ontology for his query. The *defined concept query* allows the user to define a concept based on a given shared vocabulary, which fits his understanding of a concrete

concept (Visser, 2002). In the following steps existing application ontology concepts and user-defined concepts are treated the same and will be referred to as *query concepts*.

The actual search is performed by automatically mapping between the query concept and concepts of different application ontologies within the same domain. This is possible by applying a terminological reasoner, e.g. RACER (Reasoner for A-Boxes and Concept Expressions Renamed) (Haarslev, 2001), which can work with concepts described in the Description Logic SHIQ (Horrocks, 2000). A reasoner like RACER allows the classification of data into another context by equality and subsumption. Subsumption means that if concept B satisfies the requirements for being a case of concept A, then B can automatically be classified below A (Beck, 2002). This procedure enables query processing and searching in a way that is not possible with keyword-based search.

MEASUREMENTS

```
(define-concept
ResultOfAMeasurement
  (and
    (some _quantity Quantity)
    (some _unit     Unit)
    (some _factor   Factor)
  )
)
(define-concept Measurement
  (and
    (some _has
ResultOfAMeasurement)
    (some _has Location)
    (some _has Datestamp)
    (some _has Timestamp)
  )
)
```

HYDROLOGY

```
(define-concept WaterLevel
  (and
    ResultOfAMeasurement
    (some _quantity Depth)
    (some _inspected River)
  )
)
```

Figure 2: Extract of some concepts definitions in the measurements and the hydrology domain.

QUERY CONCEPT

```
(define-concept Query_for_Measurement
  (and
    Measurement
    (some _has WaterLevel)
  )
)
```

Figure 3: Definition of the query concept for a feature type representing a water level measurement.

In our example, John can use existing domain concepts like `Measurement` and `WaterLevel` (Figure 2) to formulate his query concept `Query_for_Measurement` (Figure 3). By re-classifying this concept in RACER, it can be deduced that all three web services match the user query because they all provide measurements (having a location, a date and a time stamp) whose result is restricted to water levels. The subsumption hierarchy computed by RACER is shown in Figure 4.

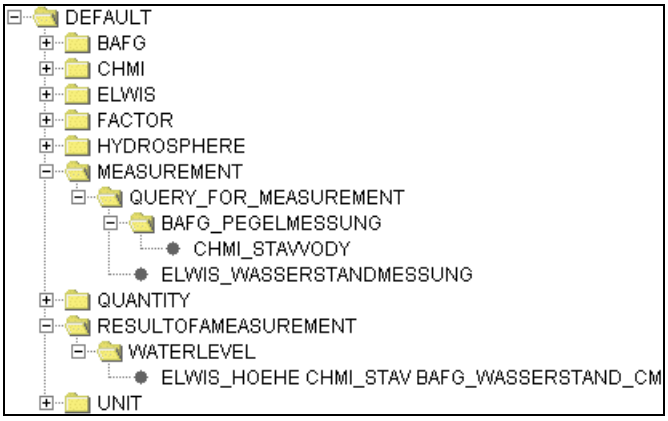

Figure 4: Subsumption hierarchy including the query concept `Query_for_Measurement`. The query concept is classified as a super-concept to all feature types provided by the three WFS.

ARCHITECTURE FOR ONTOLOGY-BASED DISCOVERY AND RETRIEVAL

The architecture we propose in this paper offers two functionalities that significantly enhance the usability of existing geographic information: using defined concept queries to overcome semantic heterogeneity problems during information discovery, and providing interpretation support for feature types and properties during information retrieval.

In order to support the advanced query capabilities described above, some new service interfaces and information items are needed in addition to the well-known components as catalogues and Web Feature Services of current SDIs. We will first describe these information items and interfaces and then sketch the information flow by means of our motivating scenario.

Components to Enable Ontology-Based Discovery and Retrieval

First, we have to provide the ontologies. For each application schema there is one application ontology that is described with the shared vocabulary of the corresponding domain. These ontologies provide the formal description of the application schema of a data source. Therefore, they are referenced from the *feature catalogue description* metadata section of the corresponding ISO 19115 documents for that data source (Figure 5). This metadata section describes content information of that data source, e.g. a list of the available feature types names.

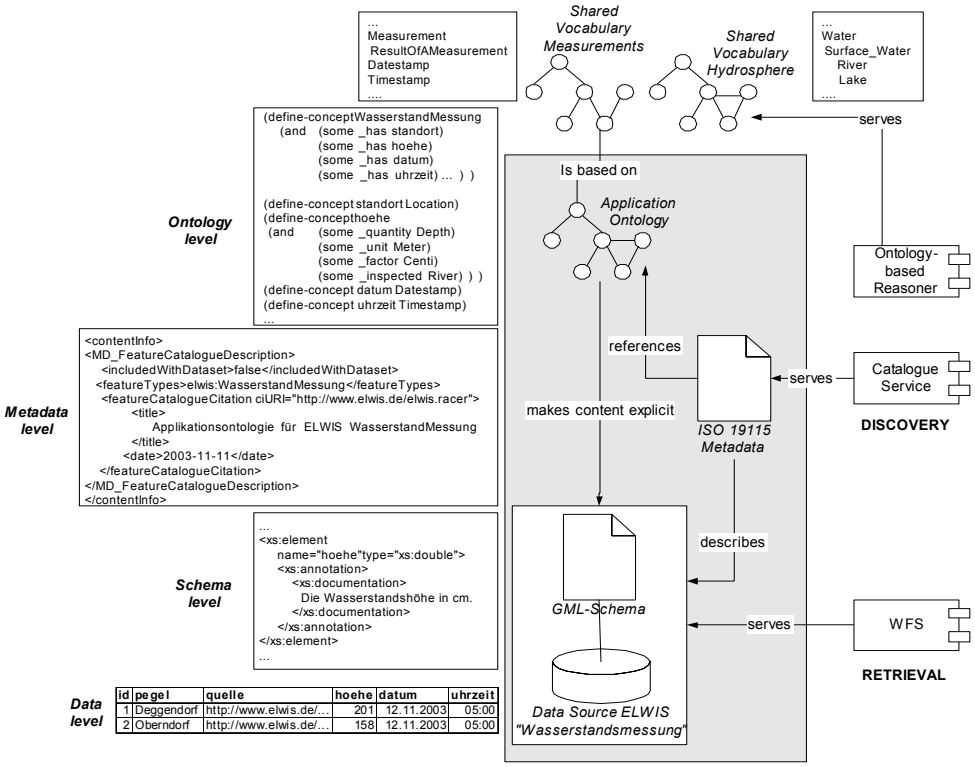

Figure 5: Services and information items required for ontology-based discovery and retrieval of geographic information.

To provide access to the ontologies, two new interfaces are defined (Figure 6): The *Concept Definition Service* interface allows access to the concepts of the shared vocabulary and application ontologies. The *Concept Query Service* interface allows to reason about possible matches with simple and defined concept search. In our prototype, both interfaces are implemented by a reasoning component that makes use of ontologies expressed in SHIQ.

The second component is a cascading catalogue service that is "aware" of the application ontologies. It provides access through the standard *OGC Stateless Catalogue Service* interface, thus implementing the decorator design pattern (Gamma, 1995). It extends the functionality of the conventional catalogue service by analysing and manipulating the filters of metadata queries. If a filter constrains a query only to return metadata results with a specific feature type in the *feature catalogue description* section, the advanced matchmaking capabilities of the *Concept Query Service* are used. The returned list of concepts is also added to the filter. This allows the decoration of any conventional standard catalogue service because the expanded filter requires only the usual exact word match. The decorating catalogue service would also enable enhanced matchmaking on other metadata elements by plugging in additional services, e.g. a gazetteer of hierarchically ordered place names.

The last component that deserves special attention is a *user interface (UI)* that utilizes the ontologies. It makes use of the *Concept Definition Service* to allow a user to formulate enhanced queries for metadata and geodata. Metadata queries for data sources with specific application schema

information are supported by allowing the construction of a query concept. The concepts from the application ontologies support the formulation of WFS queries for unknown application schemas and the interpretation of the results.

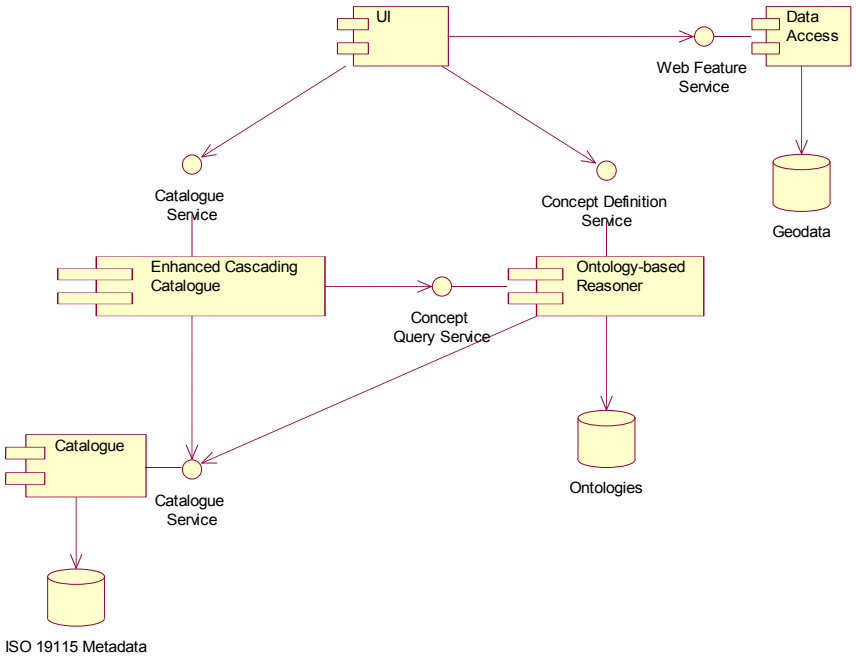

Figure 6: Components and interfaces required for ontology-based discovery and retrieval

Interaction and Information Flow in the Motivating Example

We come back to our motivating example to illustrate the interaction and information flow within the architecture (Figure 7).

John wants to construct a defined concept query in the UI using the domain's shared vocabulary. For this the *UI component* first retrieves the concepts of the shared vocabulary from the *Ontology-based Reasoner*. The user defines his query concept and a spatial query constraint that covers the Elbe catchment. The *UI component* then constructs a filter with a conjunction of the spatial constraint and a featureType constraint. For building the latter the metadata element in the *application schema information* section is constrained by the query concept.

The filter is the input of the `GetRecord` request to the *Enhanced Cascading Catalogue*. The catalogue discovers that the filter contains a constraint on the content of the data source. It uses the query concept to get a list of all matching concepts from the *Ontology-based Reasoner*. It replaces the original query concept in the filter by a disjunction of all matching concepts. This filter is forwarded to the *conventional catalogue*, which performs an exact word-based match. The results of the `GetRecord` request are finally returned to the *UI component*.

In the second step, the user wants to analyse the geodata he found with his query. The returned metadata documents contain a reference to a WFS to access the data. The data is encoded in GML. To get a description of the schema of the feature type, the *UI* invokes a `DescribeFeatureType` request and presents the GML Schema to the user. To help the user to interpret the schema, the *UI* also invokes the *Ontology-based Reasoner* to get the description of the concept defining this feature type. Because the concept is described with terms from the domain's shared vocabulary, the user can select the correct properties he is interested in, i.e. the water level and the date/time of the measurement.

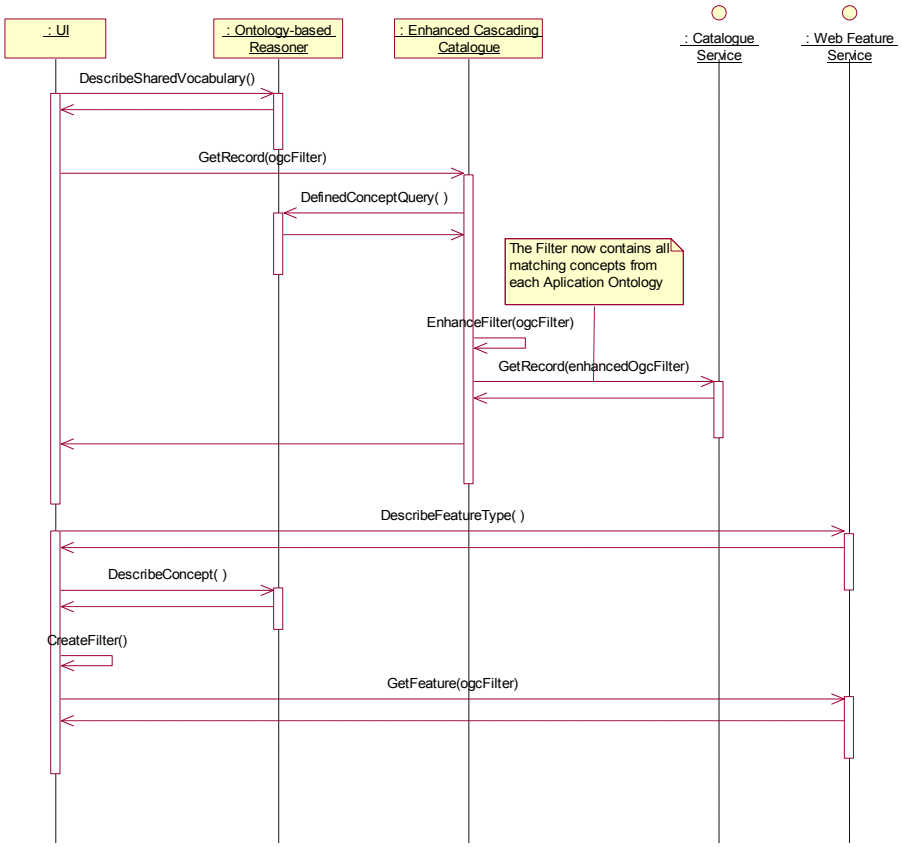

Figure 7: Information flow within the architecture for the motivating scenario.

CONCLUSION AND FUTURE WORK

We have presented an approach and architecture for ontology-based discovery and retrieval of geographic information that can contribute to solving existing problems of semantic heterogeneity. The tested scenario comprises information items with simple structures. Future tests of the architecture will include more complex application schemas and examples from other domains.

The presented architecture is component-based, i.e. it is extendable in various directions. So far, the *Enhanced Cascading Catalogue* Service and the *Reasoner* component are tightly coupled in the architecture. However, the standardized interfaces allow to extend the architecture with multiple and exchangeable components. It is also planned to extend the architecture with modules for spatial and temporal reasoning (Vögele, 2003) as well as gazetteer services.

Also, in a future version of the architecture, the tasks of discovery and retrieval will be combined in one query. The user will then be able to formulate his actual question straight away (i.e. without having to perform a query on the metadata first) using terms from the familiar shared vocabularies. The discovery and the filter formulation for retrieval will then be automated within the system. This "intelligent" query capability will enhance the usability of existing geographic information even further.

ACKNOWLEDGEMENTS

We want to thank Sören Haubrock from Delphi IMM for providing the implementation of the WFS interfaces for the use case. The work presented in this paper has been supported by the German Federal Ministry for Education and Research as part of the GEOTECHNOLOGIEN program (grant number 03F0369A). It can be referenced as publication no. GEOTECH-50.

REFERENCES

Beck, H. & H. S. Pinto (2002): Overview of Approach, Methodologies, Standards and Tools for Ontologies.

Bernstein, A. & M. Klein (2002): Towards High-Precision Service Retrieval. In Horrocks, I. & J. Hendler (eds.): The International Semantic Web Conference (Lecture Notes in Computer Science): 84-101.

Bishr, Y. (1998): Overcoming the semantic and other barriers to GIS interoperability. International Journal of Geographical Information Science 12 (4): 299-314.

Bishr, Y. & M. Radwan (2000): GDI Architectures. In Groot, R. & J. McLaughlin (eds.): Geospatial Data Infrastructures. Concepts, Cases, and Good Practice. Oxford, Oxford University Press: 135-150.

Brox, C., Y. Bishr, W. Kuhn, K. Senkler & K. Zens (2002): Toward a Geospatial Data Infrastructure for Northrhine-Westphalia. Computer, Environment and Urban Systems 26: 19-37.

Gamma, E., R. Helm, R. Johnson & J. Vlissides (1995): Design Patterns: elements of reusable object-oriented software. Boston, MA, USA, Addison-Wesley.

GDI-NRW (2002): Catalog Services für GeoDaten und GeoServices, Version 1.0. International Organization for Standardization & OpenGIS Consortium.

Gruber, T. R. [ed.] (1993): Toward Principles for the Design of Ontologies Used for Knowledge Sharing. Kluwer Academic Publishers (Formal Ontology in Conceptual Analysis and Knowledge Representation).

Haarslev, V. & R. Möller (2001): Description of the RACER System and its Applications. International Workshop on Description Logics (DL-2001).

Harvey, F., W. Kuhn, H. Pundt, Y. Bishr & C. Riedemann (1999): Semantic Interoperability: A central issue for sharing geographic information. The Annals of Regional Science 33: 213-232.

Horrocks, I., U. Sattler & S. Tobies (2000): Reasoning with Individuals for the Description Logic SHIQ. In: McAllester, D. (ed.): 17th International Conference on Automated Deduction (CADE-17) (Lecture Notes in Computer Science): 482-496.

OGC (1999): Topic 14: Semantics and Information Communities (Version 4), Open GIS Consortium.

OGC (2002): Web Feature Service Implementation Specification (OGC 02-058), Open GIS Consortium.

Schuster, G. & H. Stuckenschmidt (2001): Building shared ontologies for terminology integration. In: Stumme, G., A. Maedche & S. Staab (eds.): KI-01 Workshop on Ontologies.

Stuckenschmidt, H. (2002): Ontology-Based Information Sharing in Weakly Structured Environments. PhD Thesis, Vrije Universiteit Amsterdam: Amsterdam.

Studer, R., V. R. Benjamins & D. Fensel (1998): Knowledge Engineering: Principles and Methods. Data and Knowledge Engineering 25 (1-2): 161-197.

Uschold, M. (1998): Knowledge level modelling: concepts and terminology. The Knowledge Engineering Review 13 (1): 5-29.

Visser, U. & H. Stuckenschmidt (2002): Interoperability in GIS - Enabling Technologies. In: Ruiz, M., M. Gould & J. Ramon (eds.): 5th AGILE Conference on Geographic Information Science: 291-297.

Vögele, T., S. Hübner & G. Schuster (2003): BUSTER - An Information Broker for the Semantic Web. Künstliche Intelligenz 3 ("Semantic Web"): 31-34.

Wache, H., T. Vögele, U. Visser, H. Stuckenschmidt, G. Schuster, H. Neumann & S. Hübner (2001): Ontology-Based Integration of Information — A Survey of Existing Approaches. IJCAI-01 Workshop: Ontologies and Information Sharing: 108-117.