

# Intelligent Brokering of Environmental Information with the BUSTER System

**H. Stuckenschmidt, T. Vögele, U. Visser and R. Meyer**

Intelligent Systems Group, Center for Computing Technologies, University of Bremen. {heiner, vogele, visser, ryco}@tzi.de

***Abstract:** The World Wide Web can be seen as a huge information source. However, finding and retrieving the right data needed for an actual process is hard if not impossible. This is due to the fact that the information sources are heterogeneous and geographically distributed. An intelligent broker can help to integrate available data in the systems. In this paper we discuss the need for information sharing and propose BUSTER (Bremen University Semantic Translator for Enhanced Retrieval), an intelligent broker architecture, as a solution. In addition, we introduce a new approach for the retrieval of data with spatial information. We describe how qualitative spatial relations can help to retrieve available data. We describe a use case within the geographical domain and show how the retrieved objects can be translated semantically. We finish the paper with a demonstration of the BUSTER prototype.*

***Keywords:** Information Retrieval, Intelligent Brokering, Spatial Reasoning*

## 1 Motivation

Many application areas of information systems share the need to store and process large amounts of diverse data, which is often geographically distributed. This implies that making new data available to the system requires, that the data be transferred, into the system's specific data format. This is a process, which is very time consuming and tedious. Data acquisition, automatically or semi-automatically, often makes large-scale investment in technical infrastructure and/or manpower inevitable. These obstacles are some of the reasons behind the concept of information sharing. The solution of information sharing applies because existing information can be accessed by remote systems in order to supplement their own data basis. The advantages of successful information sharing is thus obvious for many reasons:

- Quality improvement of data due to the availability of large and complete data.
- Improvement of existing analyses and application of the new analyses.
- Cost reduction resulting from multiple use of the existing information sources.
- Avoidance of redundant data and conflicts that can arise from redundancy.

However, in order to establish efficient information sharing, difficulties arising from organizational and competence questions and many other technical problems have to be solved. Firstly, a suitable information source must be located which contains the data needed for a given task. Once the information source has been found, access to the data therein has to be provided. Furthermore, access has to be provided on a technical and informational level. In short, information sharing not only needs to provide full accessibility to the data, it also requires that the accessed data may be interpreted by the remote system. While the problem of providing access to information has been largely solved by the invention of large-scale computer networks, the problem of processing and interpreting retrieved information remains an important research topic (Visser, Stuckenschmidt et al. 2000).

## **2 The BUSTER Approach**

In systems with a large number of available data sources, it is often not trivial to find the right set of data for a given task. If, for example, an information request is submitted to an information broker, the broker has to decide which of the registered sources it should use to answer the request. The BUSTER approach addresses these three questions by providing a common interface to heterogeneous information sources in terms of an intelligent information broker (<http://www.semantic-translation.de>). A user can submit a query request to the network of integrated data sources. In this query phase several components of different levels interact.

Metadata, i.e. data describing a data source, are often used to organize and manage large collections of data sources. Typically, such metadata catalogues are based on standardized meta data formats like the Dublin Core. In the BUSTER approach, each data source is represented by a specific ontology, the so-called source ontology (Stuckenschmidt, Wache et al. 2000). It contains an explicit description of the concepts covered by the data source. Together with information about the structural and syntactic details of the data source. User queries are matched against different source ontologies. If the matching succeeds, the broker establishes a connection to the actual information source. If the matching fails,

the broker decides that there is no valuable information available and tries different information sources (Vögele, Stuckenschmidt et al. 2000)

On the syntactic level, wrappers are used to establish a communication channel to the data source(s) that have been found, that is independent of specific file formats and system implementations. Each generic wrapper covers a specific file- or data-format. For example, generic wrappers may exist for ODBC data sources, XML data files, or specific GIS formats. Still, these generic wrappers have to be configured for the specific requirements of a data source.

The mediator on the structural level uses information obtained from the wrappers and combines, integrates and abstracts them. BUSTER allows the use of different mediators which are configured by transformation rules. These rules describe in a declarative style, how the data from several sources can be integrated and transformed to the data structure of original source.

On the semantic level, we use two different tools specialized for solving the semantic heterogeneity problems based on the ontologies that describe the content of the information sources. Both tools are responsible for the context transformation, i.e. transforming data from an source-context to a goal-context. There are several ways how the context transformation can be applied. In BUSTER we consider the functional context transformation and context transformation by re-classification.

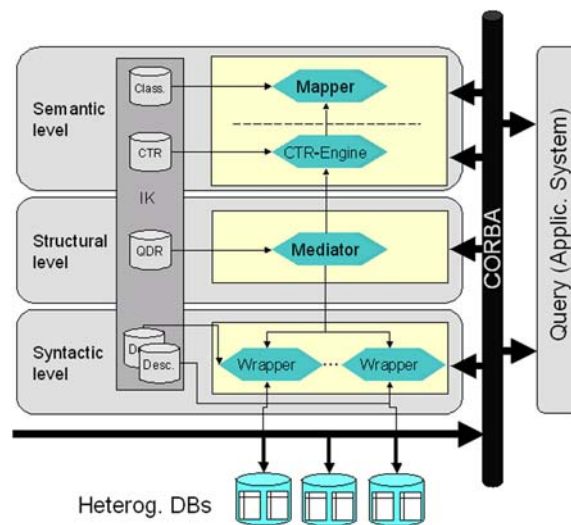


Figure 1: The BUSTER approach

BUSTER uses the language OIL for the description of content meta-data. The language has been developed in the context of the On-To-Knowledge Project

([www.ontoknowledge.org](http://www.ontoknowledge.org)) as a proposal for a language for specifying and exchanging ontologies (Fensel et al. 2000). OIL tries to provide a core set of features that have been widely accepted to be useful. OIL combines frame-based modelling primitives, reasoning facilities from Description Logics and a tight interaction with meta-data standards on the web such as RDF and XML. We used OIL to build a semantic context model of our example data by identifying a set of common properties that can be used to define a land use class.

### 3 Retrieving Spatially Related Information

In the field of environmental science, most documents and other data sources have some sort of spatial connotation. Obviously all geospatial data, i.e. data that are typically handled by GISs (Geographic Information Systems), refer to a specific geographic area. But also for non-spatial data sources, such as reports, documents and databases, references to geographic locations are typically important attributes. For example, a report about the installation of new groundwater monitoring wells very likely refers to a specific (geographic) investigation area. Consequently, spatial attributes are important for both information retrieval and the description and management of data sources with the help of metadata catalogues. However, most online systems, like metadata catalogues and other browser based information retrieval systems, offer only very little to represent and query the complex relations of data sources and their respective locations in space.

To overcome some of the shortcomings of existing approaches described in (Vögele and Stuckenschmidt 2001), we use qualitative spatial relations for information retrieval. We use the spatial configuration depicted in Figure 2 to illustrate the determination of spatial relevance on the basis of topological relations. In our example, we are concerned with different project areas in a city. The project areas are positioned within specific districts using topological relations.

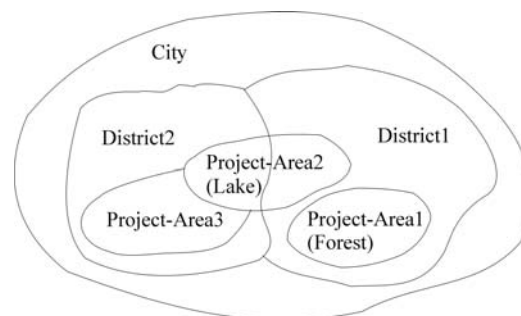


Figure 2: Spatial configuration of the example

The first relation we use to position project areas is spatial containment. *Project Area 1*, for example, is contained in *District 1*, while project *Area 3* is contained in *District 2*. We further declare that every area that is contained in another area is automatically spatially relevant with respect to the including area. This can be achieved by defining a relation `contained-in` as a special case of a relation `definitely-spatially-relevant`. Using the OIL language, we can define `contained-in` as well as its mathematical properties (i.e. transitivity) in the following way:

```
slot-def contained-in
  subslot-of definitely-spatially-relevant
  inverse contains
  properties transitive
```

We can now use the relevance relation to retrieve areas that are spatially relevant to *District 1*. Using the FaCT reasoner interface, we can formulate a query Q1 for areas spatially relevant to *District 1* in the following way:

```
(and area
  (some definitely-spatially-relevant district1))
```

Not surprisingly, the result of this query is *Project Area 1*, because it is contained in *District 1*. However, *Project Area 2* may also be of interest when querying areas related to *District 1*, because it is at least partially contained in *District 1*. We cover this kind of relevance by using another topological relation, namely partial overlap. As we are not absolutely sure that *Project Area 2* is really relevant, we use a relation `probably-spatially-relevant` to describe a weaker level of relevance. Again, we define relevance in terms of topological relations by stating, that partial overlap is a special kind of spatial relevance. The OIL definition of the relation `partially-overlapping` that is known to be symmetric is the following:

```
slot-def partially-overlapping
  subslot-of probably-spatially-relevant
  properties symmetric
```

We further define that our previous notion of relevance also falls under this new relation, because of its weaker character. The result of a query Q2 searching for areas `probably-spatially-relevant` to *District 1* consists of *Project Area 1* and *Project Area 2*, because the latter overlaps with *District 1*.

As mentioned earlier, areas in the neighborhood may also be of interest. We therefore include a further level of spatial relevance based on neighborhood defined by the relation `connected-to`. We assume that this third level of spatial relevance is even weaker than the one introduced above, because our notion of connectedness implies that there is no overlap.

```
slot-def connected-to
  subslot-of might-be-spatially-relevant
  properties symmetric
```

Using this notion of spatial relevance, we still find *Project Areas 1* and *2*. Additionally, we get *District 2* as an area spatially relevant to *District 1*. However, using OIL it is not possible to derive the spatial relevance of *Project Area 2*, which is contained in the relevant area *District 2*, in a straightforward way because we cannot chain relations in order to determine spatial relevance.

## 4 Semantic Translation of Retrieved Objects

We carried out a case study on semantic information integration based on a real life problem from the field of geographic information processing. Geographical information systems normally distinguish different types of spatial objects. Different standards exist for specifying these object types. These standards are also called catalogues. Since there is more than one standard, these catalogues compete with each other. To date, no satisfactory solution has been found to integrate these catalogues. In our evaluation we concentrate on different types of areas distinguished by the type of use.

### 4.1 Information Sources

The ATKIS catalogue (AdV 1998) is an official information system in Germany. It is a project of the head surveying offices of all the German states. The working group offers digital landscape models with different scales from 1:25.000 up to 1:1.000.000 with a detailed documentation in corresponding object catalogues. We use the large-scale catalogue OK-1000. This catalogue offers several types of objects including definitions of different types of areas. Figure 1 shows the different types of areas defined in the catalogue.

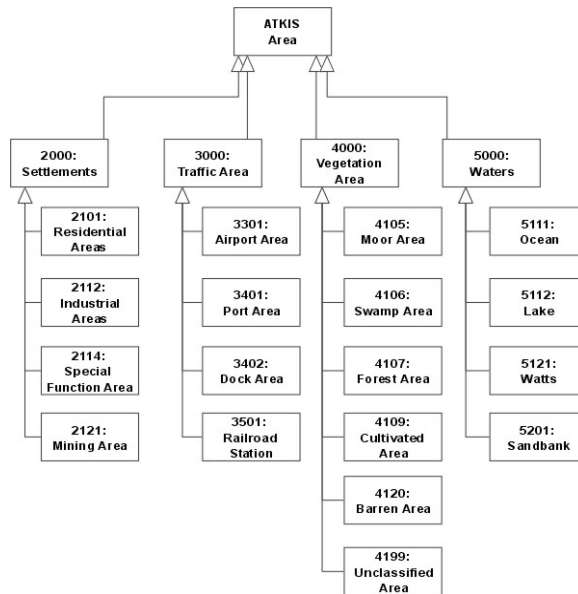


Figure 3: Class Hierarchy of the ATKIS-OK-1000 Classification

CORINE land cover (EEA 1997-1999) is a deliverable of the CORINE programme the Euro-pean Commission carried out from 1985 to 1990. The results are essentially of three types, corresponding to the three aims of the programme: (a) an information system about the state of the environment in the European Community has been created (the CORINE system). It is composed of a series of databases describing the environment in the European Community, as well as of databases with background information. (b) Nomenclatures and methodologies were developed for carrying out the programs, which are now used as the reference in the areas concerned at the community level. (c) A systematic effort was made to concert activities with all the bodies involved in the production of environmental information especially at in-ternational level. The nomenclature developed in the CORINE programme can be seen as an-other catalogue, because it also defines the taxonomy of area types (see Figure 2) with a de-scription of characteristic properties of the different land types.

The taxonomies of land-use types in figures 1 and 2 illustrate the context problem mentioned in the introduction. The set of land types chosen for these catalogues are biased by their intended use: while the ATKIS catalogue is used to administrate human activities and their impact on land use in terms of buildings and other installations, the focus of the CORINE catalogues is on the state of the environment in terms of vegetation forms.

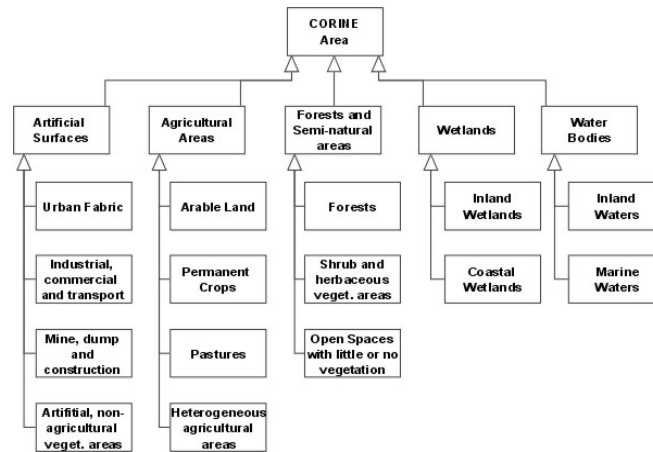


Figure 4: Class Hiererachy of the CORINE Landcover Nomenclature

Consequently, the ATKIS catalogue contains fine-grained distinctions between different types of areas used for human activities (i.e. different types of areas used for traffic and transportation) while natural areas are only distinguished very roughly. The CORINE taxonomy on the other hand contains many different kinds of natural areas (i.e. different types of cultivated areas) that are not further distinguished in the ATKIS catalogue. On the other hand, areas used for commerce and traffic are summarized in one type.

Despite of these differences in the conception of the catalogues the definition of the land-use types can be reduced to some fundamental properties. We identified six properties used to define the classes in the two catalogues. Beside *size* and *general type of use* (e.g. production, transportation or cultivation), the *kinds of structures* built on top of an area, the *shape of the ground* and *natural vegetation* as well as kinds of *cultivated plants* and three topological relations between area types are discriminating characteristics.

## 4.2 Translation Experiments

Using the definitions mentioned above, we performed a series of translation experiments, some of whose results we describe in the following. The basis for our experiment is a small CORINE landcover data set containing information about the town 'Bad Nenndorf' in Lower-Saxony. This data set is available from the German environmental agency in different formats and classifications and can therefore be used to compare and evaluate results. The data set contains areas of five different types, namely

- Discontinuous urban fabric
- Non-irrigated arable land

- Pastures
- Broad-leaved forest
- Mineral extraction site

Except for 'pastures' all these types do not directly correspond to concepts defined in our model. They are rather sub-types or special cases of the concepts we defined. Consequently, we can use the definitions from the CORINE ontology, but we have to refine the descriptions according to the additional information that is available in terms of a further specialization of the concepts.

One of the data-sets used in the case study is classified as 'broad-leaved-forest' which is a sub-class of the CORINE concept 'forest' mainly consisting of broad-leaved trees. We get a de-scription of this concept by adopting the definitions of the super-classes 'forests' and 'forests-and-semi-natural-areas' and specializing the 'has-value' constraint on the 'vegetation' slot from 'trees' to 'broad-leaved-trees'.

```

class-def broad-leaved-forest
  subclass-of area
  slot-constraint coverage
    value-type no-plants
  slot-constraint ground
    value-type land
  slot-constraint vegetation
    value-type trees OR shrubs
  has-value broad-leaved-trees

```

In this case of 'broad-leaved-forest' we got the correct result for the target hierarchy already with the first ad hoc definition of the concept to be classified. The subsumers from the target hierarchy are:

- VEGETATION-AREA
- FOREST-AREA (direct subsumer)

Looking at the target hierarchy, we can see that this is exactly the position we expected. So, we can say that at least for this case the semantic translation problem could be solved in a straightforward way using OIL and the FaCT reasoner.

## 5 Integrating Spatial and Terminological Matching

In addition to spatial representations, information about the type of place names is a central criterion for retrieval. Type information can be organized using structured concept hierarchies like thesauri and ontologies. Description logics are

very well suited for the formalization of such concept hierarchies and can be used to develop extensive ontologies of type information. Therefore, using description logics to encode both spatial relations and type information allows for the specification and fine-tuning of integrated queries.

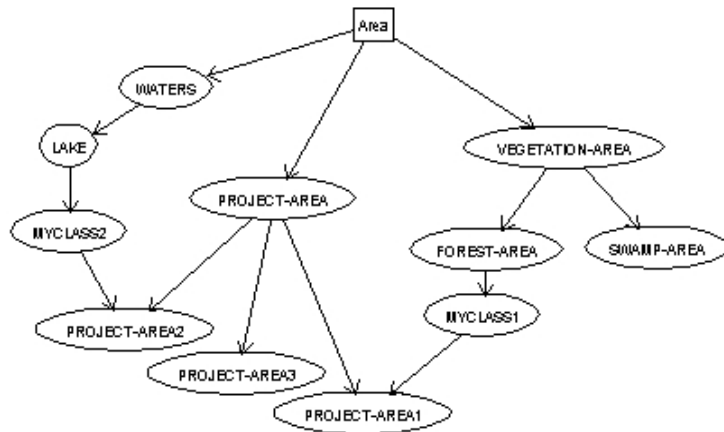


Figure 5: Sub-class relations computed by the FaCT reasoner

In order to include terminological information in queries, we further describe the project areas using class definitions and defining the areas to be instances of these classes. We might for example know that *Project Area 1* has solid ground and its vegetation consists of oaks. Using OIL we can capture this knowledge in the following class definition.

```

class-def defined MyClass1
  subclass-of Area
  slot-constraint ground
    value-type land
  slot-constraint vegetation
    has-value oak
  
```

Using the FaCT reasoner, we can automatically determine the super-class of this definition and therefore the terminological category *Project Area 1* belongs to. In our case we derive that *Project Area 1* is a “forest”, because its class definition constitutes a special case of the following general definition of a “forest area”:

```

class-def defined forest-area
  subclass-of vegetation-area
  
```

```
slot-constraint ground
  value-type land
slot-constraint vegetation
  has-value (trees or shrubs)
```

In the same way, we model the class of *Project Area 2* in such a way that it can be derived to belong to the category “lake” (see Figure for a complete class hierarchy of the example). We can use this terminological information to compute answers to more sophisticated queries. The first possibility is to restrict the type of areas we are interested in as a result. For example, we can ask for “forest areas” that might be spatially relevant to *District 1*:

```
(and forest-area
  (some might-be-spatially-relevant district1))
```

Using this additional type restricting, the result of the query reduces to *Project Area 1*, because the other areas also relevant to *District 1* are not of type forest area.

Another application of terminological information is not to seek for areas that are relevant to a specific area, but rather to a specific class of areas. For example, we can ask for areas that are spatially relevant to “lakes” in general. The corresponding query is the following:

```
(and area
  (some might-be-spatially-relevant Lake))
```

Because the logic reasoner is able to infer that *Project Area 2* is a “lake”, we retrieve all areas that are spatially related to *Project Area 2*. In our case these are *Districts 1* and *2* because they overlap with *Project Area 2* and, because of its connectedness to *Project Area 2*, also *Project Area 3*.

## 6 The BUSTER System

A first prototype of the BUSTER approach has been implemented. The current functionality includes ontology-driven search for information sources as well as schematic integration of geographical information sources. The prototype is built upon tools that have been developed at the university of Manchester to ease the use of the OIL language:

- FaCT, a logical reasoning service that can be used to check ontologies for consistency and for computing subclass relations not explicitly contained in the ontology (Horrocks 1999).
- The Ontology Editor OIEd providing a graphical interface for the definition of complex ontologies and a direct interaction with the FaCT reasoner in a client-server architecture

The editor is used to create meta-data models as well as context definitions used in the semantic translation step.

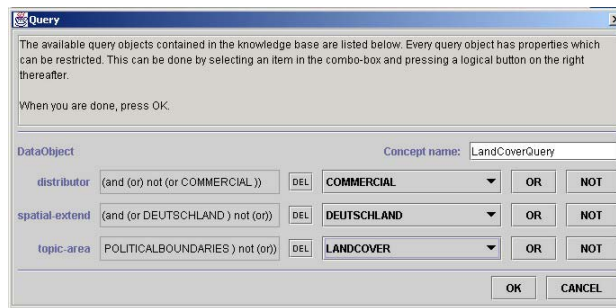


Figure 6: The BUSTER Query Interface

Figure 3 shows the system, architecture of the prototype with its query interface. The interface is dynamically created on the basis of a query model. The user is asked to restrict the defining properties of the class in order to restrict the set of all information sources to those of interest. At the moment, the FaCT reasoner is the main inference engine of the BUSTER system. The resulting class definition is passed to the reasoner that places the query in a hierarchy of classes. Each class is a surrogate for an information source. All classes placed in the subtree rooted at the query class are returned, because they fulfil the constraints defined in the query. The BUSTER system presents the information sources that matched the query to the user. 7 shows the result of a query targeting at land-use data about a special region in Lower Saxony.

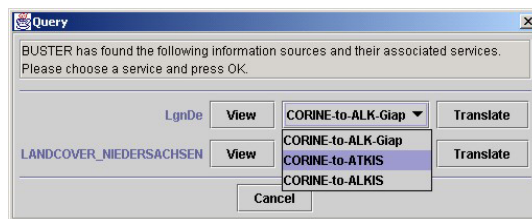


Figure 7: Results of the Query in the BUSTER Transformation Screen

The information source is named and a list of different files provided that contain different parts of the overall information. The user can now either directly view the information as a GIF image or define a target file format the information source should be converted to. Currently, in both cases the Feature Manipulation Engine FME, a conversion tool for geographical data formats is used to create the output format.

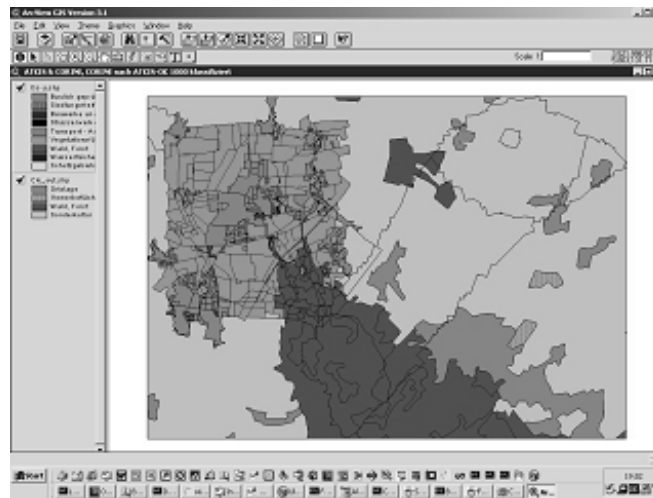


Figure 8: Integrated View on Land-Use Information

In the near future the system will be connected with the MECOTA mediator (Wache 1999), a general translation system that can be used to convert arbitrary data structures and is also capable of performing translations on the semantic level. We also aim at the integration of additional reasoning methods in addition to the FaCT reasoner in order to allow more flexible search and integration.

## Discussion

We presented the BUSTER approach for intelligent brokering of complex, spatially related information explained the knowledge-based technology underlying the approach and briefly described a prototypical implementation. The ability of the BUSTER system to combine terminological with spatial reasoning makes it an ideal platform for the exchange of environmental information which is normally related to a geographic position and frequently uses scientific vocabularies from different disciplines that need to be integrated when searching

for a special piece of information. At the moment the system is still in the development phase, but first experiences have been made that show that the approach can be successfully applied in principle.

## Literature

- ADV, 1998, Amtliches Topographisch Kartographisches Informationssystem Atkis. Technical report, Landesvermessungs-amt NRW, Bonn.
- EEA, 1997-1999, Corine land cover. Technical Guide. Technical report, European Environmental Agency. ETC/LC, Euro-peanTopic Centre on Land Cover.
- Fensel, D., I. Horrocks, et al. (2000). OIL in a Nutshell. 12th International Conference on Knowledge Engineering and Knowledge Management EKAW 2000, Juan-les-Pins, France.
- Horrocks, I. (1999). FaCT and iFaCT. Proceedings of the International Workshop on Description Logics (DL'99). P. Lambrix, A. Borgida, M. Lenzerini, R. Möller and P. Patel-Schneider: 133-135.
- Stuckenschmidt, H., H. Wache, et al. (2000). Enabling Technologies for Interoperability. Workshop on the 14th International Symposium of Computer Science for Environmental Protection, Bonn, Germany, TZI, University of Bremen.
- Visser, U., H. Stuckenschmidt, et al. (2000). Using Environmental Information Efficiently: Sharing Data and Knowledge from Heterogeneous Sources. Environmental Information Systems in Industry and Public Administration. C. Rautenstrauch. Hershey, PA, IDEA Group.
- Vögele, T., Stuckenschmidt, H, et al. (2000). Towards Intelligent Brokering of Geoinformation. Urban and Rural Data Management (UDMS), Delft, Electronic.
- Vögele, T, Stuckenschmidt, H (2001) Enhancing Gezeteeers with Qualitative Spatial Concepts. In Proceedings of the Workshop on Hypermedia in Environmental Protection, Ulm.
- Wache, H. (1999). Towards Rule-Based Context Transformation in Mediators. International Workshop on Engineering Federated Information Systems (EFIS 99), Kùhlungsborn, Germany, Infix-Verlag.