Spatio-Terminological Information Retrieval using the BUSTER System

Ubbo Visser and Thomas Vögele

TZI, Center for Computing Technologies,
University of Bremen,
Universitätsallee 21-23
D-28359 Bremen, Germany
e-mail: {visser|vogele}@tzi.de

1 Introduction

Over the last decade several approaches with regard to intelligent information integration have been proposed (e.g. Ontobroker, PICSEL, BUSTER; [Wache et al., 2001]). Some of these systems use ontologies for integration. Lately, this approach has been widely accepted [Grüninger and Uschold, 2002]. The Bremen University Semantic Translator for Enhanced Retrieval (BUSTER), a middleware based also on ontologies, has been developed at the Center for Computing Technologies. This system consists of two parts: BUSTER/Q, a tool for intelligent information retrieval, and BUSTER/SI, a tool for the semantic integration of heterogeneous data sources. Both tools can be integrated into an infrastructure of distributed geoservices such as the GDI NRW [Kuhn et al., 2000], see also figure 1. The focus of this paper will be BUSTER/Q, the information retrieval service.

BUSTER/Q offers a tool to manage and select geodata sources and geoservices suitable for specific tasks and purposes. To be able to fulfil this task, BUSTER/Q has to have access to detailed descriptions of the data sources and services. Such detailed metadata are made in the form of Comprehensive Source Descriptions (CSD), located at the site of the data source or geoservice, and formalized in XML/RDF format. Through an automatic retrieval mechanism, the individual CSDs are collected in regular time intervals and integrated into the BUSTER/Q CSD database.

Based on the metadata provided by the CSDs and appropriate qualitative terminological and spatial models (spatial ontologies), BUSTER/Q supports integrated queries of the type *concept@location*. In the following, we present a short description of the main features of a CSD, as well as of terminological and spatial queries.

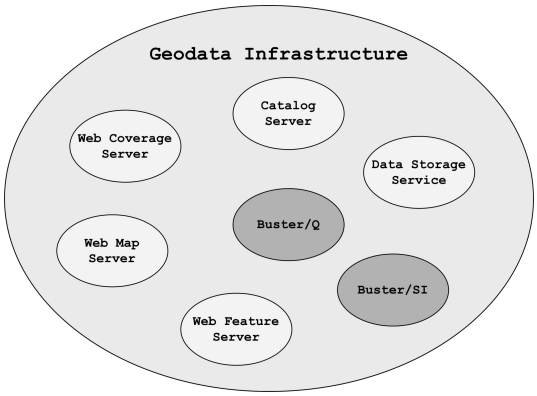

Fig. 1. Distributed geodata infrastructure

2 The Comprehensive Source Description

In order to describe existing data metadata have to be used. Hence, we have to find an eligible language for the description. Over the last decade numerous meta data formats have emerged (e.g. Dublin Core, ISO/TC211). A good overview about existing meta information systems can be found in [Visser et al., 2001]. Since we are not dependent on any specific domain, in fact we would like to use a general way to describe the data, we decided to use the Dublin Core Element Set, version 1.1 as a de facto basis for our CSD. The definitions utilize a formal standard for the description of metadata elements. The authors claim that the formalization helps to improve consistency with other metadata communities and enhances the clarity, scope, and internal consistency of the Dublin Core metadata element definitions.

However, some of the given elements are not sophisticated enough in their expressivity (e.g. relation element) or lack formal semantics (e.g. description element). Thus, there is a need for additional qualifiers for those elements, which are described in a language that provides formal semantics (e.g. DAML, OIL, SHIQ). We can use this kind of description logics to encode additional features. We use the RDF(S) syntax if possible to ensure a wide acceptance with respect to accessibility and usability. Please note that the expressiveness of RDF(S) is sometimes not enough. We then refer to explicit ontologies available on the WWW. The following elements are refined for our CSD:

- Coverage: Since there is no further distinction between spatial and temporal coverage, this element has to be refined.
  - Spatial: The recommended best practice from DCMI is to select a value from a controlled vocabulary and that, where appropriate, named places or time periods be used in preference to numeric identifiers such as sets of coordinates or date ranges. Examples are *DCMI Point* to describe a point in space using its geographic coordinates, *ISO 3166* a code for the representation of names of countries, *DCMI Box* that identifies a region of space using its geographic limits. The last recommendation is *TGN*, the GETTY Thesaurus of Geographic Names (see http://shiva.pub.getty.edu/tgn_browser/). We decided on the latter.
  - Temporal: The recommend best practice here is to use one of the two following encoding schemes: *DCMI Period*, a specification of the limits of a time interval, and *W3C-DTF*, the W3C encoding rules for dates and times - a profile based on ISO 8601 (see also: http://www.w3.org/TR/NOTE-datetime). We use the latter since the main reason to have this CSD is to describe information sources on the WWW.
- Description: Description may include but is not limited to: an abstract, table of contents, reference to a graphical representation of content or a free-text account of the content. The semantics of this kind of representation are limited with regards to machine readable meaning of the content. Hence, we restrict the description to a formal description logic, namely DAML+OIL or SHIQ. The vocabulary used to describe this A-Boxes has to be one of the vocabularies used in the "relation" element.
- Relation: The qualifiers that refine the relation element as recommended by DCMI is limited. Therefore, we need to extend these qualifiers by references that also point to ontologies, gazetteers or thesauries. A relation is described as a XML name space describing the URI of the corresponding vocabulary and a prefix to mark terms from this vocabulary.
- Subject: The qualifiers recommended by DCMI for the subject element contain common lists of keyword from various sources (e.g. the Library of Congress Subject Headings, Medical Subject Headings, Universal Decimal Classification). In BUSTER, we use the subject element accordingly, it remains a list of significant keywords to describe the information source but the keywords have to be chosen from a controlled vocabulary referred by the relation element.
- Rights: Despite the intellectual property rights we also have to consider access rights for special user groups. In the moment, there is no further specification.

The final paper will describe the 15 elements of the Dublin Core Metadata Format with the above mentioned additional features. A concrete example will be given to show the reader how a comprehensive description of an information source would look like.

3 Terminological Queries

Once an information source has been annotated with all the information needed, complex queries can be made to the BUSTER/Q system. BUSTER/Q is based on terminological ontologies that have been modelled in advance. The system can be connected to two logical reasoner available on the WWW, namely the FaCT system by the University of Manchester [Horrocks, 1999] and the RACER system by the University of Hamburg [Haarslev and Möller, 2001].

The user gets to choose from several domain ontologies (e.g. an environmental domain or an electronic devices domain) and can then specify queries in a common vocabulary provided by BUSTER/Q, which is underlying the given ontologies. One example is to describe an electronic device which is flexible and the material is PVC. The query is then processed by the reasoner and the result would contain a list of all the information sources which contain the specified target concept, an installation pipe in this case.

In a first step, all the information sources listed in a CSD database would be checked with respect to their relationship with this particular domain. A CSD, for example could describe an information source that contains data classified by the ETIM (etim.de) or ECLASS (eclass.de) catalogue (two common catalogue systems which are used in industry). The second step is the following: the query is processed by the logical reasoning systems, which are able to match the query against the given concept. This means that a given query (a description of a target concept) can be matched against various and different! catalogue systems. The result is a list of eligible information sources containing the target concepts. The target concepts could vary according to the underlying catalogue systems.

4 Spatial Queries

A user-friendly and, from a cognitive perspective, sound method to specify spatial queries as well as to index data sources and services is the use of placenames. Placenames are typically organized in gazetteers [Hill, 2000] [Riekert, 1999]. Schlieder et al. [Schlieder et al., 2001,**?**] propose an extension to gazetteers in the form of placename structures based on qualitative spatial models. These models, or spatial ontologies, use graph representations of hierarchically organized polygonal tessellations as a basis to reason about the spatial relevance of one placename with respect to another.

In BUSTER/Q, the user is able to select a specific spatial ontology, based on which he can specify a spatial query. By selecting a placename, the user defines the target area of the spatial query. Using the selected spatial ontology, the spatial reasoner integrated in BUSTER/Q evaluates the query and computes a list of placenames that are spatially relevant to the target placename. Spatial relevance is computed based on a combined evaluation of partonomic and neighborhood relations between placenames. The result is a weighted list of spatially relevant placenames.

5 Combined Spatio-Terminological Queries

BUSTER/Q combines both lists, the list of relevant concepts, and the list of spatially relevant placenames, into one database query. This database query is applied to the BUSTER CSD database. The result is a weighted list of data sources and geoservices matching both the terminological and the spatial query.

The final paper will provide a detailed view on the CSD and on both the terminological and spatial reasoning. We will give an example of how the combined queries look like and discuss the benefits and also the shortcomings of our approach. We are developing a prototype of this approach at the moment and would be able to give a live demonstration at the conference.

References

[Grüninger and Uschold, 2002] Grüninger, M. and Uschold, M. (2002). Ontologies and semantic integration. In *Software Agents for the Warfighter*. Institute for Human and Machine Cognition (IHMC), University of West Florida. in preparation.

[Haarslev and Möller, 2001] Haarslev, V. and Möller, R. (2001). High performance reasoning with very large knowledge bases. In Nebel, B., editor, *International Joint Conferences on Artificial Intelligence (IJCAI)*, volume 1, pages 161–166, Seattle, WA. Morgan Kaufman.

[Hill, 2000] Hill, L. L. (2000). Core elements of digital gazetteers: placenames, categories, and footprints. In Borbinha, J. and Baker, T., editors, *ECDL 2000*, Research and Advanced Technology for Digital Libraries, pages 280–290, Lisbon, Portugal.

[Horrocks, 1999] Horrocks, I. (1999). Fact and ifact. In Lambrix, P., Borgida, A., Lenzerini, M., ller, R. M. ö., and Patel-Schneider, P., editors, *Proceedings of the International Workshop on Description Logics (DL'99)*, pages 133–135.

[Kuhn et al., 2000] Kuhn, W., Basedow, S., Brox, C., Riedemann, C., Rossol, H., Senkler, K., and Zens, K. (2000). Geospatial data infrastructure (gdi) north-rhine westfalia - reference model 3.0. Technical report, Geoinformatik Münster.

[Riekert, 1999] Riekert, W.-F. (1999). Erschließung von fachinformationen im internet mit hilfe von thesauri und gazetteers. In Dade, C. and Schulz, B., editors, *Management von Umweltinformationen in vernetzten Umgebungen, 2nd workshop HMI*, Nürnberg.

[Schlieder et al., 2001] Schlieder, C., Vögele, T., and Visser, U. (2001). Qualitative spatial reasoning for information retrieval by gazetteers. In *COSIT'01*, Morro Bay.

[Schlieder and Vögele, 2002] Schlieder, C. and Vögele, T. (2002). Indexing and browsing digital maps with intelligent thumbnails. In *accepted at Spatial Data Handling 2002 (SDH'02)*, Ottawa, Canada. Springer.

[Visser et al., 2001] Visser, U., Stuckenschmidt, H., Wache, H., and Vögele, T. (2001). Using environmental information efficiently: Sharing data and knowledge from heterogeneous sources. In Rautenstrauch, C. and Patig, S., editors, *Environmental Information Systems in Industry and Public Administration*, pages 41–73. IDEA Group, Hershey, USA & London, UK.

[Wache et al., 2001] Wache, H., Vögele, T., Visser, U., Stuckenschmidt, H., Schuster, G., Neumann, H., and Hübner, S. (2001). Ontology-based integration of information - a survey of existing approaches. In Pérez, A. G., Grüninger, M., Stuckenschmidt, H., and Uschold, M., editors, *IJCAI-01 Workshop: Ontologies and Information Sharing*, pages 108–117, Seattle, WA.