

Overcoming Differences of Meaning during the Discovery and Retrieval of Geospatial Information in Spatial Data Infrastructures

Michael Lutz, Eva Klien

Institute for Geoinformatics, University of Münster
Robert-Koch-Str. 26-28, 48149 Münster, Germany
{m.lutz, klien}@uni-muenster.de

Advances in sensor technology are revolutionizing the way that geospatial information is collected and analyzed. Already today (and even more so in the foreseeable future) sensors provide continuous streams of geospatial information (GI). While this opens opportunities for improving the way decisions are taken in a variety of fields, it also presents a number of challenges. We are focusing on the crucial question of *how to find suitable information* for a given task (GI discovery) and *how to access it* once it has been found (GI retrieval).

Spatial Data Infrastructures (SDIs) provide searchable catalogs of information descriptions (metadata) in a standardized format (ISO/TC-211 2003) and through a standardized interface (OGC 2004). For GI retrieval, too, standardized interfaces are provided and widely used in the form of Web Feature Services (OGC 2002). However, a number of problems are caused by differences in meaning (semantic heterogeneity) during both GI discovery and retrieval.

Current GI discovery is largely based on string-matching keywords or other search terms. Even though natural language processing techniques can increase the semantic relevance of search results w.r.t. to the search request (e.g. Richardson & Smeaton 1995), keyword-based techniques are inherently restricted by the ambiguities of natural language. As a result, keyword-based search can have low recall if different terminology is used and/or low precision if terms are homonymous or because of their limited possibilities to express complex queries (Bernstein & Klein 2002).

Once a suitable information source has been discovered and could be accessed through a WFS interface, ambiguity still presents a problem. While it is possible to obtain syntactic descriptions of the application schema of the feature types, this description is often not sufficient for interpreting the meaning of its attributes. This makes it difficult (or in some cases impossible) to create a query expression that actually retrieves the desired information.

To overcome these problems, we propose to use ontological (i.e. explicit, formal and widely-agreed) descriptions of information sources. Thus, the semantics of the content of information sources become machine-interpretable, and users are enabled to pose concise and expressive queries. Furthermore, logical reasoning can be used to discover implicit relationships between search terms and information descriptions as well as to flexibly construct taxonomies for classifying information sources (Klien *et al.* 2004).

In recent years, a number of similar approaches have been proposed (e.g. Lin *et al.* 2003). However, they have not been widely adopted outside academia yet. We believe that a number of challenges need to be addressed before ontological approaches will reach widespread adoption.

- *Integration into existing SDI architecture.* To facilitate adoption, it should be possible to easily integrate ontology-based approaches for GI discovery and retrieval into existing SDI architectures. This is preferably realized by extending existing components and adding new components where necessary. For example, in (Klien *et al.* 2004) we present a component that understands queries that contain ontological query concepts. It accesses a reasoning component to enrich the query with additional ontology concepts and then sends the enriched query to a conventional catalog service. As the “enhanced” catalog service reuses the standardized catalog service interface it can also deal with normal (non-semantic) queries.
- *Standardization of domain vocabularies.* Ontology-based approaches overcome some of the problems of conventional metadata standards by enabling expressive but flexible descriptions. However, standardization is still necessary at the level of the used domain vocabulary in order to enable

interoperability *between ontological descriptions*. In the geospatial domain, a body of rich knowledge models already exists in the form of ISO TC 211 and OGC standards, which could be used as the basis for developing an already widely agreed domain vocabulary. Unfortunately, this knowledge is only semi-formal (as text or UML models). How to extract and represent this knowledge in a formal way is therefore an important research issue (Probst *et al.* 2004).

- *User support for creating semantic descriptions of geographic information.* The ontology-based approach of GI discovery and retrieval is only feasible, if all information resources registered in the SDI are semantically described. Taking into account how difficult it is to get data and service providers to specify *conventional* metadata, user support for generating *semantic* descriptions is definitely a crucial issue. This includes methods for automatically extracting (parts of) these semantic descriptions as well as intuitive user interfaces that hide the complexity of ontology languages from the provider.
- *Hiding ontologies and reasoning from the user.* Similarly, SDI users need to be supported in formulating their queries. As understanding and using ontology languages is beyond normal users, an intuitive query language and/or graphical user interface should allow a requester to intuitively formulate a query using a well-known domain vocabulary. As a first step in this direction, we are currently developing a prototype, which provides the requester with a GIS-like query interface as well as a SQL-like query language.
- *Support (semi-)automatic composition of complex processing services.* Finally, if Spatial Webs are also to include processing capabilities, it should be possible to seamlessly combine several data sources and processing services in order to provide a user with an answer to his actual question. Ideally, the composition should go unnoticed by the user. This vision would require methods for semantically describing services and for using these descriptions for service discovery (Lutz 2004). Also, it should be possible to automatically generate mediators based on the semantic descriptions that bridge heterogeneities between the services within such a composite service.

References

- Bernstein, A. & Klein, M. (2002): *Towards High-Precision Service Retrieval*, in: Horrocks, I. & J. Hendler (ed.): *The Semantic Web - First International Semantic Web Conference (ISWC 2002)*: 84-101.
- ISO/TC-211 (2003): *ISO 19115:2003. Geographic information - Metadata*, International Organization for Standardization.
- Klien, E., Einspanier, U., Lutz, M. & Hübner, S. (2004): *An Architecture for Ontology-Based Discovery and Retrieval of Geographic Information*, in: Toppen, F. & P. Prastacos (ed.): *7th Conference on Geographic Information Science (AGILE 2004)*: 179-188.
- Lin, K., Ludäscher, B., Brodaric, B., Seber, D., Baru, C. & Sinha, K. A. (2003): *Semantic Mediation Services in Geologic Data Integration: A Case Study from the GEON Grid*, Geological Society of America (GSA) Annual Meeting 35 (6).
- Lutz, M. (2004): *Non-taxonomic Relations in Semantic Service Discovery and Composition*, 1st "Ontology in Action" Workshop, in conjunction with 16th Conference on Software Engineering and Knowledge Engineering (SEKE 2004): 482-485.
- OGC (2002): *Web Feature Service Implementation Specification, Version 1.0.0 (OGC Implementation Specification 02-058)*, Open Geospatial Consortium.
- OGC (2004): *Catalogue Services Specification, Version 2.0 (OGC Implementation Specification 04-021r2)*, Open Geospatial Consortium.
- Probst, F., Gibotti, F., Pazos, A., Esbri, M. A., Benigno, M., Gutiérrez, M. & Kuhn, W. (2004): *Connecting ISO and OGC Models to the Semantic Web*, 3rd International Conference on Geographic Information Science.
- Richardson, R. & Smeaton, A. F. (1995): *Using WordNet in a Knowledge-based Approach to Information Retrieval (Technical Report CA-0395)*, Dublin City University.