

SPATIAL DATA ACCESSIBILITY AND THE SEMANTIC WEB

Femke Reitsma, University of Maryland College Park, MD (femke@geog.umd.edu)
Mentor: Lola Olsen, NASA GSFC, Code 902

Abstract. *As the volume of spatial data increases and the number of spatial data web portals proliferates, new approaches to searching across multiple sites are needed to enhance knowledge discovery. The Semantic Web provides such an opportunity for automated retrieval of spatial data, presenting a general solution for standard access and querying for spatial data. The advantages of applying the concepts of the Semantic Web are explored in the first steps towards an implementation using NASA's Global Change Master Directory's (GCMD) metadata holdings.*

INTRODUCTION

The quantity and quality of spatial data is continually increasing as our ability to observe the world grows through advances in technology, such as the development of new satellites and sensor networks. This is particularly evident in the realms of Earth System Science, where petabytes of satellite imagery are captured, stored, and ready for distribution. As the volume of data grows, increasingly sophisticated methods of data retrieval are needed as a basis for knowledge discovery. This paper considers the potential of the Semantic Web for enhancing spatial data distribution and describes an application of Semantic Web technology for metadata distribution. The value of the Semantic Web to database retrieval is described in the next section. Progress through an exploratory implementation of spatial data representation on the Semantic Web with NASA's Global Change Master Directory (GCMD) follows in Section 3. Future enhancements are addressed in the final section.

Value of Semantically Enhanced Databases

The proliferation of spatial data providers often makes it difficult to find the information needed, particularly with increasingly interdisciplinary approaches to research. Engaging any one of these distribution facilities, let alone the multitude of available

services through web portals, requires a substantial amount of time for familiarization with the methods of data extraction presented (for human or computational agent interaction). Attempts at providing a one-stop shop have yet to prove their effectiveness in data delivery, for example, the National Spatial Data Infrastructure's Geospatial One-Stop (<http://www.geo-one-stop.gov/>).

The data providers typically adhere to standards that define how the data must be described, such as U.S. federally mandated Federal Geographic Data Committee's (FGDC) Content Standard on Digital Geospatial Metadata (CSDGM) (see <http://fgdc.er.usgs.gov/metadata/contstan.html>) and ISO 19115 Geographic Information Metadata standard (<http://www.isotc211.org/scope.htm#19115>). Standards for accessing spatial data have not been well applied for data in general. In particular, the client/server model of ISO 23950 (Information Retrieval - Z39.50, <http://www.loc.gov/z3950/agency/>) does not provide a general solution (Evans 1997). Furthermore, standards for data exchange, such as GML (<http://opengis.net/gml/01-029/GML2.html>), provide no semantic content.

Data needs to be easily machine accessible through a standard method of querying across web portals. In addition tools need to be developed that can easily query these in concert. The Semantic Web, in essence, provides this default standard for

interoperability, whereby the expression of content that is semantically annotated in a web ontology language (RDF, RDFS, or OWL) can then be queried by autonomous agents in the same manner regardless of the database structure. As described by Berners-Lee (Berners-Lee 1998), “one of the main driving forces for the Semantic Web, has always been the expression, on the Web, of the vast amount of relational database information in a way that can be processed by machines”.

Semantic query languages, one of the core being RQL

(<http://139.91.183.30:9090/RDF/RQL/>), provide the means to access RDF (and potentially OWL) descriptions with a minimal knowledge of the schema(s) employed (Karvounarakis et al. 2000). The flexibility of semantic query languages allows one to traverse the graph through a predefined set of steps, whereby one can express types of queries that cannot be expressed in existing languages with schema querying capabilities (see Karvounarakis et al. 2000 for examples). In developing a data format that is flexible enough to represent any form of domain knowledge, the Semantic Web can potentially query any form of information (Decker 2002).

Uploading the GCMD to the Semantic Web

The Global Change Master Directory (GCMD) is currently in the process of exploring the potential of the Semantic Web for distributing its metadata holdings. As described below, we have converted a subset of metadata files to RDF/OWL, converted the keywords to ontologies, and tested queries over these files. Results are not yet available online, as software testing continues and the automation of some processes remains.

Overview of the GCMD

The GCMD provides metadata for Earth science data sets and services that are relevant to global change research, including metadata covering areas of research such as

atmosphere, biosphere, hydrosphere, geology and geophysics, and human dimensions of global change. The metadata is currently provided through a web portal (<http://gcmd.nasa.gov/>) in the Directory Interchange Format (DIF), a de-facto standard used to create directory entries which describe data. (The DIF standard is FGDC and ISO 19115 compatible). Querying these data relies on a set of controlled keywords that define the bounds of knowledge discovery above that of a free text search (available at: <http://gcmd.nasa.gov/Resources/valids/index.html>). There is also a free text search available, which is enhanced with spatial and temporal qualifiers. For automated querying and large scale data retrieval, an API is provided.

IMPLEMENTATION

Three key steps have been implemented in providing Semantic Web access to the GCMD’s metadata. First, the ontologies and DIF schema were developed. An XSLT style sheet was then created to convert the DIF files to RDF (Clark 1999). Finally, a software environment was selected and tested that would support the distribution of the GCMD’s metadata content in a Semantic Web language.

In the development of ontologies for the GCMD, it was important to use the existing intellectual capital embodied in the controlled vocabularies (Qin and Paling 2001), while recognizing that some changes are necessary in order to match the data model embodied in Semantic Web languages. For example, the following hierarchy is taken from the science keywords:

EARTH SCIENCE > Hydrosphere > Ground Water > Saltwater Intrusion

Earth Science is a discipline, whereas Hydrosphere, Ground Water, and Saltwater Intrusion form a class hierarchy of topics that

are studied by that discipline. Thus, the mapping between the keywords and the ontologies required careful consideration of the appropriate meaning of the terms and the relationship among those terms, which needed to be elicited from those who defined them. For example, the Variable Fetch is a measurable property of the Term Ocean Waves; however, the Variable Fisheries is a sub-topic of the Term Agricultural Aquatic Sciences.

Beyond the keywords, a spatial ontology was derived from an ESRI spatial dataset (that distributed with the ArcGIS software). This was converted to RDF, where continents, countries, and regional administrative units were related by an `isPartOf` property. Unfortunately, the `owl:inverseOf` property defined in OWL could not be used in describing the inverse of `isPartOf` as spatial containment, because software does not yet support this level of expressivity. The ontologies were also linked to other ontologies, such as the Dublin Core (<http://dublincore.org/>) and Cyc (<http://www.cyc.com/cyc-2-1/intro-public.html>) ontologies, in order to avoid the creation of a monolithic structure that does not tap into the potential of the Semantic Web.

The RDF DIF schema was created by directly mapping the DIF to RDF (for an example of a DIF see:

http://gcmd.nasa.gov/User/difguide/annot_template.html). This schema formed the basis of an XSLT style sheet, which was used to automatically convert XML versions of the DIF metadata files to RDF and OWL. This conversion incorporated both the DIF Schema and the ontologies, mapping not only the XML tags but also the GCMD's keywords within those tags. It was found to be unnecessary to encode the full "CATEGORY > TOPIC > TERM > VARIABLE" hierarchy of the science keywords to the RDF DIF as this was defined in the ontology. Rather, only the instances of finest granularity were converted, as all other levels could be inferred

through the `rdfs:Class/rdfs:subclassOf` and `rdf:Property/rdfs:subPropertyOf` relationships.

Sesame was selected as the software environment to distribute the GCMD's metadata on the Semantic Web

(Broekstra et al. 2003, also see:

<http://sesame.aidadministrator.nl/>).

Sesame is a storage and querying middleware for the Semantic Web, which will be used by the GCMD to provide its metadata content in RDF (Figure 1). Thus far Sesame has been tested with PostgreSQL, an object-relational DBMS. An RDF DIF subset, the ontologies, and the RDF DIF Schema were all loaded and successfully queried.

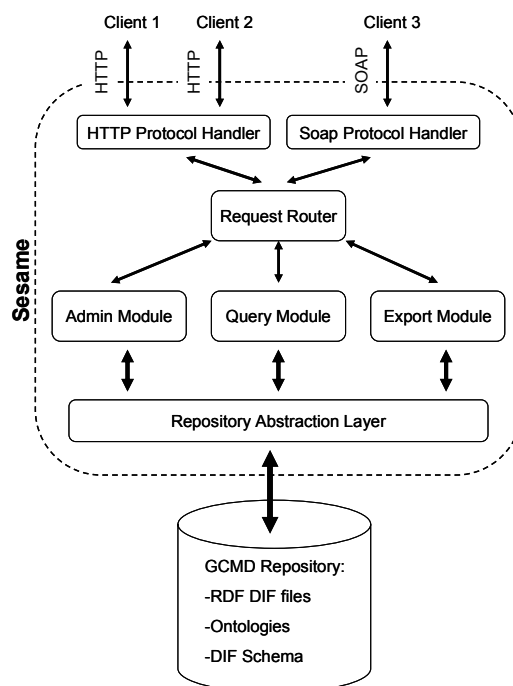


Figure 1: Architecture of GCMD implementation with Sesame (modified from Broekstra et al. (2003))

OUTLOOK

The next steps for the GCMD involve: enriching the ontologies with more content, thereby supporting more comprehensive querying; creating a method to automatically update the ontology with any change in the keywords; and uploading the developments thus far to an experimental web site where users can access the DIF files in RDF and query them.

Challenges that are evident on the near horizon include: the problem of ontology change, which involves versioning and maintenance (Decker 2002); the development of a GUI that allows for much easier interaction with the semantically marked-up DIF files; and extending the language of the Semantic Web to incorporate spatial relationships. Furthermore, given the nascent nature of technology supporting the Semantic Web, issues remain regarding the coordination and integration of middleware products that are only now in their infancy (some will succeed; others will fail). Remaining at the cutting edge of software development while maintaining an operational system is critical for the GCMD's future developments with the Semantic Web.

ACKNOWLEDGEMENTS

We gratefully acknowledge the assistance of Professor Jim Hendler and the Mindswap group (<http://owl.mindswap.org/>).

REFERENCES

- Berners-Lee, T. (1998). Relational Databases on the Semantic Web. See <http://www.w3.org/DesignIssues/RDB-RDF.html>
- Broekstra, J., A. Kampman and F. van Harmelen (2003). Sesame: A Generic Architecture for Storing and Querying RDF and RDF Schema. In: Towards the Semantic Web: Ontology-driven Knowledge Management. J. Davies, D. Fensel and F. van Harmelen. Chichester, John Wiley & Sons Ltd: 71-89.
- Clark, J. (1999). XSL Transformations (XSL-T), W3C Recommendation, 1999. See <http://www.w3.org/tr/xslt>
- Decker, S. (2002). Semantic Web and Databases: Relationships and Some Open Problems. In Proceedings of the NSF-EU Workshop on Database and Information Systems Research for Semantic Web and Enterprises, Amicalola Falls and State Park, Georgia, USA. See <http://lsdis.cs.uga.edu/SemNSF/SemWeb-DBIS-Workshop-Proc.pdf>
- Evans, P. (1997). Z39.50 Stand-alone Client Software Review. Biblio Tech Review. See <http://www.biblio-tech.com>
- Karvounarakis, G., V. Christophides, D. Plexousakis and S. Alexaki (2000). Querying Community Web Portals. Technical report, Institute of Computer Science, FORTH, Heraklion, Greece. See <http://www.ics.forth.gr/proj/isst/RDF/RQL/rql.pdf>
- Qin, J. and S. Paling (2001). Converting a Controlled Vocabulary into an Ontology: the Case of GEM. Information Research 6(2). Available at: <http://InformationR.net/ir/6-2/paper94.html>