An Ontological Approach to Geoscience Dataset Cataloging

Blumenthal, M.B., del Corral, J., Bell, M., and Grover-Kopec, E.

International Research Institute for Climate and Society (IRI), Earth Institute, Columbia University
Lamont-Doherty Campus, 61 Route 9W, Palisades, NY 10964
benno@iri.columbia.edu, jdcorral@iri.columbia.edu, mbell@iri.columbia.edu, grover@iri.columbia.edu
http://iridl.ldeo.columbia.edu

Abstract

Geoscience has traditionally struggled with metadata for datasets. There are many fixed metadata schemas to facilitate the search and retrieval of datasets but they have trouble representing the richness of datasets coming from an ever-evolving science. For the IRI Data Library of datasets, we want a more relational and extensible paradigm for metadata. The RDF/OWL framework in the form of OWL ontologies presents that possibility.

Ontology Development

The IRI ontology development (see fig. 1) for cataloging datasets began with an OWL schema creation consisting of classes, subclasses and properties. This schema incorporated both original and imported classes and properties. We imported existing ontologies from Dublin Core (dc), Really Simple Syndication (rss), part of NASA's Global Change Master Directory (GCMD), the Climate and Forecast Metadata (CF), and NASA's Semantic Web for Earth and Environmental Terminology (SWEET). Both GCMD and CF ontologies are hosted and prepared by the Marine Metadata Interoperability Project (MMI). We used Protege as our ontology creation editor.

Populating the ontology with individuals (sometimes referred to as instances) was the next step. First we converted existing and inferred dataset attributes into RDF/OWL metadata. Some hand editing with Protege was needed for information that was not available in our original system. We added SWRL rules and SeRQL CONSTRUCTs to be able to generate inferred triples.

A perl crawler using Redland was used to gather and store the information for the ontology. The resulting triples were then transferred to Sesame where the RDFS and OWL semantics were evaluated by Storage and Inference Layers (SAILs) and stored in a Sesame RDF database.

SWRL rules were translated into SeRQL CONSTRUCTs. Sesame was used to process these CONSTRUCTs and their inferences. We then stored the virtual triples with the asserted triples.

Iterative Queries

The search interface implements a separate search ontology that includes taxa, facets, and properties to represent the desired faceted search choices. This search ontology is connected to the dataset description ontologies by the RDF/OWL framework. The interface uses SeRQL queries to evaluate user-selected criteria in the Sesame RDF database. The interface returns rss links to the datasets and a subset of the original search terms that describe the datasets selected.

Datasets and the terms that describe them are connected by the property isDescribedBy. Terms are interrelated by directlyImplies properties, and datasets are interrelated by Dublin Core isPartOf and isReplacedBy properties (these properties are used to reduce repetition in the search returns). Subsequent selection criteria narrow the range of possible datasets by creating new SeRQL queries that operate on the remaining possible datasets.

Display Query Results

We used perl and HTML to present the user with a web-based interface to the search ontology facets and terms. As the user makes a facet-term selection, the web display is updated to reflect the remaining possible choices and the datasets that match the selection criteria. A list of criteria that have already been selected is maintained to allow the user to de-select a criterion and see the resulting range of datasets available.

Conclusion

Ontologies can be used to formulate highly adaptive cataloging and search mechanisms. The ability to import existing ontologies encourages the reuse of schemas developed by experts in their specific fields. Machine interpreted knowledge representations like RDF/OWL ontologies open the door for widely distributed knowledge bases.

Future Plans

We hope to expand our use of SAILs within Sesame. We plan to use these to reduce the number of inference processing steps with our ontologies. We will begin testing Sesame 2.0-alpha as a migration path from our current use of Sesame 1.2, which will allow us to move our persistent store into Sesame. We plan to use other metadata ontologies from the geosciences to more fully describe our datasets, and to be able to describe them with standardized terms. We plan to use our ontologies to represent and deliver our dataset metadata in conformance with various geoscience metadata standards, such as Directory Interchange Format (DIF), Federal Geographic Data Committee (FGDC), and the International Organization for Standardization (ISO).

References

Daconta, M. C., Obrst, L. J. and Smith, K. T., 2003, *The Semantic Web*. Indianapolis, Indiana: Wiley Publishing.

Aduna B.V., Sirma Al Ltd., 2002-2006, *User Guide for Sesame*, Free Software Foundation.

Noy, N. and McGuinness, D. L., 2001, *Ontology Development 101*, Palo Alto, California: Stanford Medical Informatics.

Beckett, D., 2002-2006, *Redland RDF Application Framework*, University of Bristol.

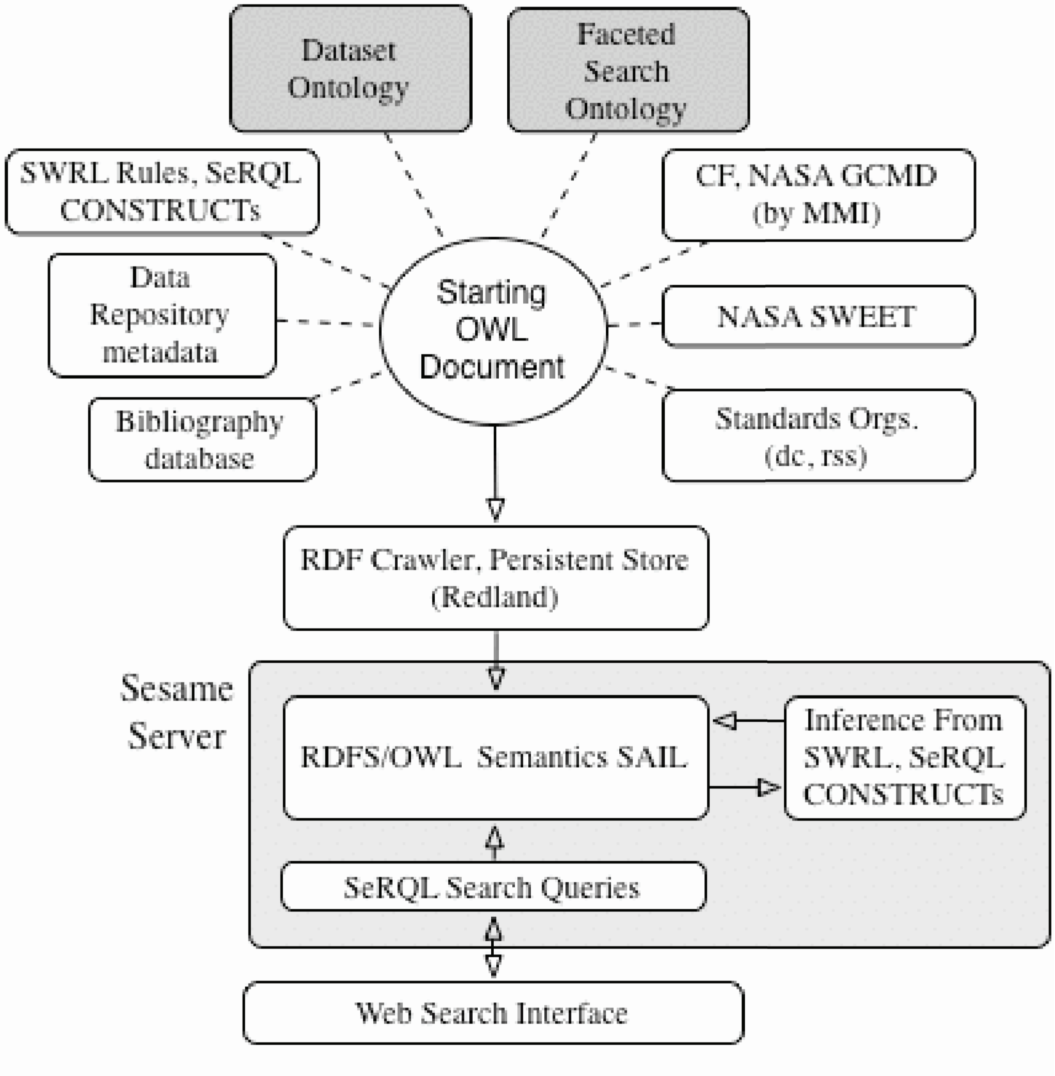

Figure 1 IRI Dataset Ontology System Architecture