Semantic Search Integration to Climate Data

Ranjeet Devarakonda, Giriprakash Palanisamy, Line C. Pouchard, Biva Shrestha
Oak Ridge National Laboratory, 1 Bethel Valley Road, Oak Ridge, TN 37831
devarakondar@ornl.gov, palanisamyg@ornl.gov, pouchardlc@ornl.gov, shresthab@ornl.gov

POSTER EXTENDED ABSTRACT

In this paper we present how research projects at Oak Ridge National Laboratory are using Semantic Search capabilities to help scientists perform their research. We will discuss how the Mercury metadata search system, with the help of the semantic search capability, is being used to find, retrieve, and link climate change data.

INTRODUCTION

The Semantic web is changing the nature of how information is produced and shared. In addition to human-human interactions, machine-machine interactions are becoming more and more common. Structured data, as in machine-readable data, are the key to enabling these interactions. This is especially true for scientific data with large, diverse, multi-disciplinary data volumes, therefore making data integration a necessity. Much of these data do not live in the cloud or on the Web, but rather in multi-institutional data centers that provide tools and add value through quality assurance, validation, creation, dissemination, and analysis of data.

Mercury [1], developed at Oak Ridge National Laboratory, is a metadata search, discovery and access system. It provides access to more than 200,000 biogeo-chemical and ecological records and is used by approximately 30,000 scientists each month. With the breadth of sciences represented within the Mercury metadata records, scientists can potentially address key interdisciplinary scientific challenges related to climate change including carbon sequestration, advance of seasons, anticipated environmental and ecological impacts, as well as address questions related to the mitigation of these potential effects. However, the wealth of data and metadata also makes it difficult to pinpoint the datasets relevant to particular scientific inquiries. We have applied semantic technologies — ontologies, in particular — to improve the relevance of metadata search results. In this paper, we focus on the use of ontologies—formal machine-readable descriptions of the domain—to facilitate data search and discovery via Mercury.

IMPLEMENTATION

Mercury currently provides access to over 200,000 metadata records. It supports several widely used metadata standards and protocols such as the Federal Geographic Data Committee, Dublin Core, Darwin Core, the Ecological Metadata Language, the International Standards Organization's ISO-19115, Extensible Mark-up Language (XML), and Library of Congress protocols Z39.50 and Search/Retrieve via URL.

The Mercury architecture (Figure 1) includes a harvester, an indexing tool, and a user interface. Mercury's harvester typically harvests metadata records from publicly available external servers. Data providers and principal investigators create metadata for their datasets and place these metadata files in a publicly accessible place such as a web directory or FTP directory. Mercury then harvests these metadata files, builds the centralized index, and makes the index available for the Mercury search user interface. Mercury also harvests metadata records from external catalogs using the Open Archives Initiative Protocol for Metadata Harvest (OAI-PMH) [2] and other web-based harvesting techniques.

We have integrated the Semantic Web for Earth and Environmental Terminology (SWEET) Ontology [3] with the Mercury search interface. SWEET is a mature foundational ontology developed at the NASA Jet Propulsion Laboratory. In order to incorporate the SWEET ontology relevant to earth sciences into the Mercury architecture, we chose BioPortal as an ontology repository. BioPortal is a community - based ontology repository developed by the National Center for Biomedical Ontology (NCBO) [4]. The BioPortal at Oak Ridge National Laboratory's Distributed Active Archive Center (ORNL DAAC) allows users to browse ontologies and to look for specific ontologies that have terms relevant for their work. The mappings between ontologies in BioPortal not only allow users to compare the use of related terms in different ontologies, but also allow analysis of how whole ontologies compare with one another. BioPortal provides access to the ontologies through a REST interface, thus enabling easy integration with Mercury. In order to provide access to ontology entities in the ORNL DAAC BioPortal instance, we designed an ontology service that allows integration of ontology entities into search results. The Mercury search

978-1-4799-5158-1/14/$31.00 ©2014 IEEE

system passes search parameters to BioPortal, which renders one or several entities (classes, properties, terms) through the REST interface. The user can select any of these entities as additional search parameters for Mercury or directly expose the results shown by the ontology sub - class terms.

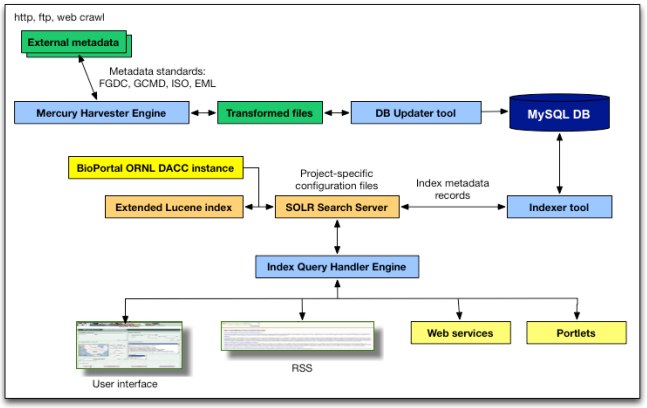

Figure 1. Mercury Search Engine architecture and its integration with the BioPortal ORNL DAAC instance. The Mercury Search service calls the BioPortal Instance to add ontology Knowledge to the queries.

The Mercury interface already uses a faceted search approach to display search results. Mercury, in addition to querying the local index, passes the query parameter to the BioPortal search API as a Java *document* and receives the response as a *list*. The *list* contains three facets (Concepts, Super-classes and Sub-classes) provided by the Ontology's REST response. We extract the facet values from the response (Figure 2) and integrate them with the existing Mercury facets (Figure 3).

```
Document doc = builder
      .build("http://localhost:8280/Mercury3_MN2/mn/xml/SWEET")
  List beanList = beans.selectNodes(doc);
  Iterator i = beanList.iterator();
while (i.hasNext()) {
  String[] row = new String[2];
  Element bean = (Element) i.next();
  String s1 = bean.getChild("conceptIdShort").getTextTrim();
  String s2 = bean.getChild("preferredName").getTextTrim();
  row[0] = s1.substring(s1.lastIndexOf("/") + 1).trim();
  row[1] = cs.splitCamelCase(s2);
  als_main.add(row);
}
```

Figure 2. Querying BioPortal

Example: The ontology service exposes "humus" as an extra search term for Mercury in the first discovery session about "biomass."

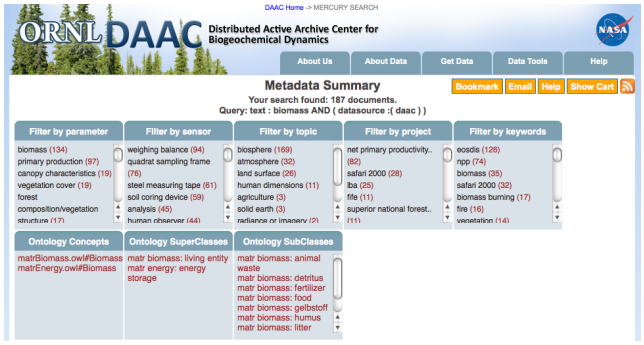

Figure 3. Faceted results of a Semantic search on "biomass"

Humus is a sub-class of biomass in SWEET. Five additional datasets are returned by a search on "humus", that were not returned by a search on "biomass". "Biomass" also acquires scientific context when the ontology service exposes that it can be a form of Energy Storage and a Living Entity.

ACKNOWLEDGMENTS

Oak Ridge National Laboratory is managed by the UT-Battelle, LLC, for the U.S. Department of Energy under contract DE-AC05-00OR22725.

We thank Thomas A. Boden from Environmental Sciences Division, Oak Ridge National Laboratory for his useful comments and language editing which have greatly improved the manuscript.

REFERENCES

[1] R. Devarakonda, et al., Mercury: reusable metadata management, data discovery and access system. Earth Science Informatics, 2010. 3(1 - 2): p. 87 - 94.

[2] R. Devarakonda, et al., Data sharing and retrieval using OAI-PMH. Earth Science Informatics, 2011. 4(1): p. 1-5.

[3] R.G. Raskin. and Pan, M.J., Knowledge representation in the semantic web for Earth and environmental terminology (SWEET). Computers & Geosciences, 2005. 31(9): p. 1119 - 1125.

[4] P.L. Whetzel, Noy, N.F., Shah, N.H., Alexander, P.R., Nyulas, C.I., Tudorache, T., and Musen, M.A., BioPortal: Enhanced Functionality via New Web services from the National Center for Biomedical Ontology to Access and Use Ontologies in Software Applications. Nucleic Acids Research (NAR), 2011. 39(Web Server issue): p.W541-5