

A Linked Science Investigation: Enhancing Climate Change Data Discovery with Ontologies and Semantic Technologies

Line C. POUCHARD^a, Ranjeet DEVARAKONDA^b, Marcia BRANSTETTER^b, and Natasha NOY^c

^a*Purdue University Libraries, West Lafayette, IN 47906*

^b*Oak Ridge National Laboratory, 1 Bethel Valley Road, Oak Ridge, TN 37831*

^c*Google*

Abstract. *Linked Science* is the practice of inter-connecting scientific assets by publishing, sharing and linking scientific data and processes in end-to-end loosely coupled workflows that allow the sharing and re-use of scientific data. Much of this data does 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, curation, dissemination, and analysis of the data. In this paper, we make the case for the use of scientific scenarios in Linked Science. We propose a scenario in river-channel transport that requires biogeochemical experimental data and global climate-simulation model data from many sources. We focus on the use of ontologies—formal machine-readable descriptions of the domain—to facilitate search and discovery of this data. Mercury, developed at Oak Ridge National Laboratory, is a tool for distributed metadata harvesting, search and retrieval. Mercury currently provides uniform access to more than 100,000 metadata records; 30,000 scientists use it each month. We augmented search in Mercury with ontologies, such as the ontologies in the Semantic Web for Earth and Environmental Terminology (SWEET) collection by prototyping a component that provides access to the ontology terms from Mercury. We evaluate the coverage of SWEET for the ORNL Distributed Active Archive Center (ORNL DAAC).

Keywords: Linked Science, ontologies, BioPortal, semantic search, climate change, data discovery

1. Introduction

The ways in which scientists conduct research in earth sciences, chemistry, biology, geography, ecology, sociology, and other scientific fields is changing significantly. Often, the most challenging research questions require them to understand and use practices, data, methods and software from many scientific disciplines. *Linked Science* is the practice of inter-connecting scientific assets by publishing, sharing and linking scientific data and processes in end-to-end loosely coupled workflows that allow the sharing and re-use of scientific data [2-4]. Linked Science requires new ways of integrating and aggregating structured and unstructured data and information derived from physical, chemical, biological, sociological, and other traditional fields of scientific study. Linked Science is grounded in interdisciplinary research and highlights the reasons why such research is arduous: a scientist who is already an expert in a domain must become fluent in the language and practices of another domain in order to start addressing a scientific question. Semantic technologies such as ontologies can help with this by providing annotations and descriptions of the concepts and relationships in a domain of science.

Ontologies define the concepts in a domain of discourse, provide constraints on the values, and define formal semantics that enable knowledge representation and automated reasoning. The World-Wide Web Consortium (W3C) has defined OWL, a formal language for representing and sharing ontologies on the Web, enabling scientists to publish and integrate metadata using standard Web protocols. One can think of an ontology as a taxonomy of terms representing the concepts of a domain with added rules and relationships that can be used by computer algorithms. Ontologies and semantic descriptions of the scientific data and processes provide the necessary objects supporting the production of new knowledge by allowing interoperability of the processes, shared annotations and integration of the data.

This chapter is organized as follows. The next section describes requirements and examples in Linked Science and motivates the use of ontologies in Linked Science. Section 3 describes two widely used ontologies in Earth and Environmental Sciences. In Section 4, we describe the domain application and provide a scientific scenario where heterogeneous data sources must be collected to perform a scientific investigation of climate change for river water transport. We characterize the various types of datasets available in this domain. In section 5 we focus on the design and implementation of our system and the integration of several open-source components to improve data discovery using semantics. We also describe a prototype tool highlighting the use of ontologies developed for this scenario. In Section 6, we present the results that we obtained with the prototype and we evaluate the ontology coverage. Section 7 describes related work. In the final section we analyze our results and then present our conclusion and future directions.

2. Linked Science: requirements and examples

Some examples of Linked Science, such as DataONE, put a special emphasis on working from scientific scenarios. Data Observation Network for Earth (DataONE) is the foundation of new innovative environmental science through a distributed framework and sustainable cyber-infrastructure that meets the needs of science and society for open, persistent, robust, and secure access to well-described and easily discovered Earth observational data [5]. The goal of DataONE is to ensure the preservation, access, use and reuse of multi-scale, multi-discipline, and multi-national science data via cyber-infrastructure elements and a broad education and outreach program. DataONE researchers have developed a number of scientific scenarios that require such multi-disciplinary integration of data. In DataONE's initial efforts at using data to address scientific research projects, scientists used bird observations and a variety of environmental data layers to estimate changes in the occurrence of bird species seasonally in the conterminous U.S [6, 7].

Gil and colleagues [8] propose another Linked Science example and developed a scenario focused on understanding the carbon cycle in water that requires integrating data and analyses by scientists studying river, lake, ocean, and coastal ecosystems. Here, a semantic framework allows collective metadata editing and acknowledgement of scientists' contributions around the scenarios. The Semantic Water Quality Portal [9], also an example of Linked Science, integrates domain data related to water quality from several agencies, including the

US Geological Survey (USGS), the Environmental Protection Agency (EPA) and multiple regulation ontologies. Ontologies enable detection of pollution events and communities can monitor pollution results according to a regulation of their choice.

Linked Science relies upon the collection, organization, classification, storage, discovery, access, transport, distribution, sub-setting, aggregation, dissemination, and visualization of large, diverse types of data. Scientists store and disseminate data through archives and data centers, supported by organizations in government, academia, and industry. These data centers guarantee data quality and reproducible transformation of data through processes so that the credibility of scientific results is preserved. This is essential in the study of climate change, where results influence national and international policy. Data centers include experimental, observational, and computer-generated data. They provide tools and add value through quality assurance, validation, curation, dissemination, and analysis of the data. These data typically cannot be consumed by a browser, an audio or video reader, and usually require specialized applications that these data centers also provide.

Discovery and access to this data poses a major challenge, one that we describe below. On the Web, with Linked Open Data, every resource has a unique identifier. By contrast, with Linked Science, uniform access to datasets provided by unique identifiers is not always available because the data does not live in the cloud or on the Web, but in multi-institutional data centers. Uniform data access would be advantageous to searches and for discovering new links between datasets and/or scientists. Globally unique identifiers, a unique reference number used as an identifier in software and on the web, like a Uniform Resource Identifier, are gaining some traction but not universally used. Common schemes like the Digital Object Identifiers system, a unique reference ISO standard, are often linked to publications rather than datasets. Critically, each dataset must be accompanied by discovery metadata to enable access. Metadata must specify which services or software can consume the data and where they are offered to allow automation of processes. In addition, metadata must describe the way that the data was generated, potential errors, and uncertainty or variability in the calculations and measurements.

Thus, not only must we have a vocabulary to describe this extensive metadata, but also this vocabulary needs to be shared among multiple data providers and we must be able to perform automated reasoning in order to discover, access, and integrate data described by this metadata. We investigate the use of *formal ontologies* to represent metadata vocabularies. Ontologies help with data access in Linked Science as they can provide additional keywords for a search.

3. Ontologies in the Earth and Environmental Sciences

3.1. The Semantic Web for Earth and Environmental Terminology (SWEET)

Many ontologies exist in the Earth and Environmental Sciences domain. The Semantic Web for Earth and Environmental Terminology (SWEET) [10] is one of the most widely used and the most extensive. It is a set of ontologies that includes more than four thousands five hundred classes of terms and related concepts in Earth and Space science. The SWEET ontologies were developed according to the principles of scalability, application independence, natural language independence, orthogonality, and community involvement [10]. Scalability implies that the ontologies are extensible, that is, the concepts in an ontology can be re-used and further specified to represent domain or sub-domain knowledge. Application independence guarantees that the ontologies can be used regardless of the original intent or implementation. The SWEET ontologies are implemented using OWL, the Web Ontology Language, so that any application that supports OWL can use these ontologies, access the terms representing concepts, and take advantage of the properties and relationships between the terms for inferring new knowledge. SWEET is independent from natural language as the knowledge is embedded into concepts rather than terms. However, the concepts are represented in natural language that provides labels for the concepts. In SWEET, composite labels such as “ice cap” are presented using camel case, which could lead to additional processing requirements when one attempts to use SWEET with natural language applications. Orthogonality implies that compound concepts are decomposed into component parts. Orthogonality ensures that re-use of concepts is more easily achieved as the concepts are reduced following the principle of reductionism, by which

specialists decompose entities into their component parts [10]. Community involvement is assured by the Earth Science Information Partners (ESIP) federation which has the governance of SWEET. The current stable version at the time of this writing is version 2.3.

These concepts in SWEET are divided into 3 integrative ontologies and 9 faceted ontologies representing orthogonal dimensions (Figure 1). Each box represents a separate ontology, and the connecting lines indicate where major properties are used to define concepts [1].

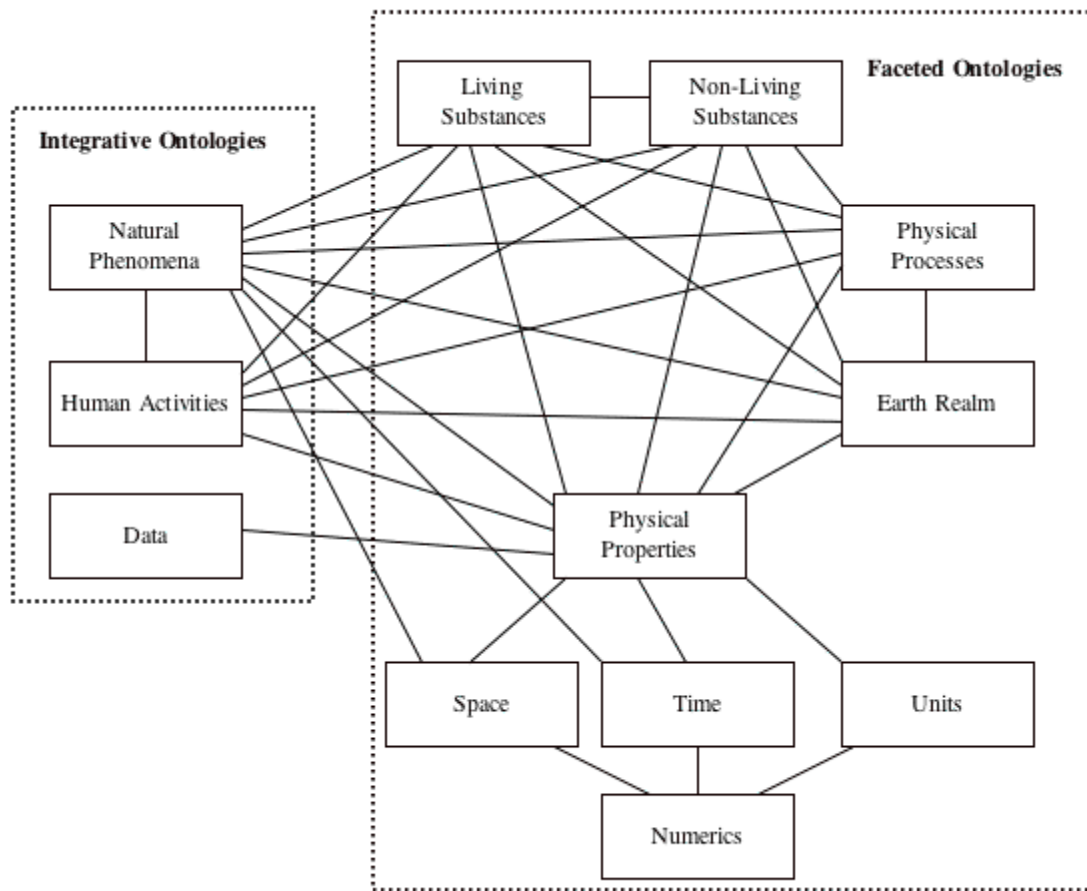


Figure 1: SWEET ontologies and their relationships [1].

The SWEET ontologies are currently under the governance of the Earth Science Information Partners, a broad-based information community of data and technology practitioners working together on interoperability efforts across Earth and Environmental sciences. SWEET 2.3 currently contains over 4,500 classes organized in 200 OWL ontologies classifying 9 top-level classes. For SWEET 2.3 these top-level classes are:

- Representation (math, space, science, time, data)
- Realm (ocean, land surface, terrestrial hydrosphere, atmosphere, heliosphere, cryosphere, geosphere)
- Phenomena (macro-scale ecological and physical)

- Processes (micro-scale physical, biological, chemical, and mathematical)
- Matter (living thing, material thing)
- Human Activities (decision, commerce, jurisdiction, environmental, research),
- Property (binary, categorical, ordinal, quantity)
- Role (biological, chemical, geographic, impact, representative, trust),
- Relation (human, physical, space, time, chemical).

SWEET is a mature ontology that can provide search terms for our investigation. For example a search on “carbon” in SWEET returns “Carbon sequestration,” “carbon footprint,” “Dissolved organic carbon concentration protocol” and “kartz,” that all can be used for new searches. We used the SWEET ontologies in our tool development.

3.2. The Environment Ontology

The product of a community effort, the Environment Ontology (ENVO)[11] contextualizes biomedical and biological entities by describing the environment in which samples are taken. Motivated by the need to describe the environmental origins of tissue and pathogen, the concepts in this ontology also apply to the description of sample environments in the Earth Sciences. ENVO contains 1397 classes and is formalized in OWL and the OBO format¹. Efforts have been made to align ENVO’s four top-level classes with the Basic Formal Ontology (BFO), an upper ontology that provides a semantic foundation for many domain ontologies. The four top-level classes include: Environmental System (Biome and Habitat), Environmental Feature, Environmental Condition, and Environmental Material.

In the alignment with BFO, the Environmental System is a subclass of BFO’s system that includes a material entity within its site and causally influences that entity. This class is under development and its definition subject to change at the time of this writing. The Biome is a major class of ecologically similar communities of plants, animals, and other organisms. Biomes are often defined in terms of plant structure, leaf types, plant spacing and climate. Biomes are not defined by genetic, taxonomic, or historical similarities, but are often identified with patterns of ecological succession and climax vegetation. The Biome class is divided into terrestrial biome and aquatic biome. The Habitat class, a spatial region having environmental qualities that may sustain an organism or a community of organisms, is also undergoing revision at this time. The Environmental Feature class refers to environment types that are described by a single entity with a strong causal influence on its surrounding space, for instance, a coral reef. In contrast with biomes, environments of this type make no specific reference to the ecological communities or populations they support. Three sub-classes are currently defined: geographic feature, organic feature and mesoscopic physical object, a type of smaller scale, discrete, solid, transportable environment such as a carcass. The Environmental Condition defines specific ranges of determinate qualities (e.g. a temperature range, a solar irradiation range). The sub-classes of Environmental Conditions include arid, polar, sub-polar, sub-tropical, temperate, and tropical. These sub-classes do not refer to geographic ranges but to the qualities indicated above. They may be used as conditional properties to specify a biome. For instance, the Temperate Broadleaf Forest Biome has condition temperate. The Environmental Material class refers to masses, volumes, or portions of some medium included in an environmental system. It is understood to be more complex than a simple collection of material entities. For instance, the material “soil” contains aggregates of rock particles, plants, fungi, microbes, water and airspace.

The ENVO ontology is still very much under development and alignment with BFO through a community process. Early adoption by the metagenomics community led to ENVO’s acceptance as a project within the framework of the Genomic Standards Consortium. The broader ENVO consortium has developed through workshops, meetings and user engagement with participants and domain experts from a wide range of domains, including biodiversity, biomedicine, marine ecology, microbiology, and ethno-geography. ENVO’s primary

¹ <http://www.obofoundry.org>

² <http://www.earthsystemgrid.org/>

usage is to provide terms for annotations in the description of biological samples. We did not use it to provide search terms for our prototype.

In the next section, we propose a scientific scenario in climate change that illustrates Linked Science and presents the challenges for obtaining heterogeneous datasets.

4. Use case scenario and datasets

Consider the following scenario (Figure 2). A hydrologist focuses on validating model simulation multi-decadal trends for nutrient transport in a river channel within a watershed. In this case we are considering the Community Climate System Model [12]. This climate model simulation of the earth system used to investigate climate change has four components: land, sea ice, ocean and atmosphere. The land component currently includes simulations of river flow. Future models of the earth system will contain biogeochemical species such as nitrogen, carbon, and phosphorus compounds (e.g., those contained in fertilizers). Changes in the chemistry of rivers from two different scenarios are particularly relevant to climate change. First, biogeochemical species resulting from fertilizer use are washed from the soil, carried from water streams into larger rivers, and eventually end up in coastal oceans. Second, deforestation from biomass burning also causes changes to the chemical composition of the water that flows into rivers. The transport of biogeochemical species, particularly riverine nitrogen, may have an even larger effect: these species cause hypoxia (reduction in the oxygen concentration in water) and fish mortality in the coastal oceans [13]. In order to characterize these effects realistically, the hydrologist will need access to two types of data, which are generally available to earth scientists: (1) *computational* data that record the results of computer modeling and simulation; and (2) *observational* data that contain results of specific measurements.

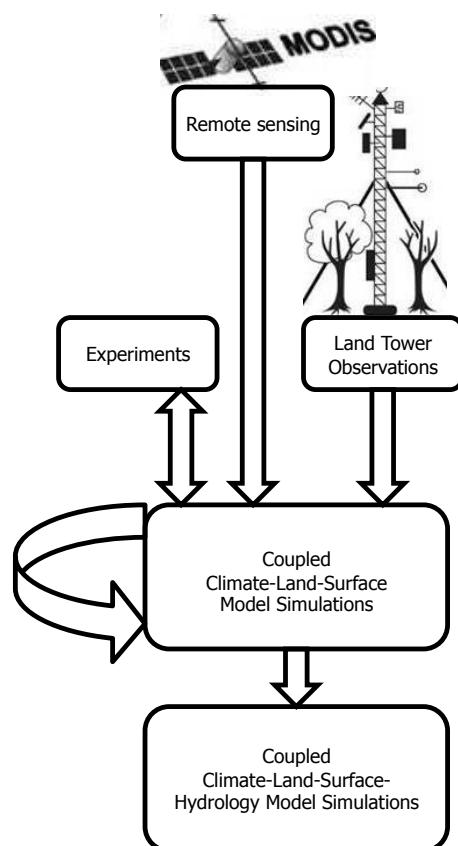


Figure 2 Different types of data in our river channel transport use case scenario

Getting the data

In our use case, the computational data will include models of river flow and transport of biogeochemical species; the observational data will describe stream flow, water quality, precipitation, air and water temperature, sediment data, biogeochemical species, and soil moisture. For computational-model data, our hydrologist can turn to the Earth System Grid Federation (ESGF²) gateway at the National Center for Atmospheric Research [14]. At the time of this writing, it contains 3,384 datasets of computational data totaling about 1.3 Petabytes of data and representing 368 variables. She will need to know, however, that file names in this source attempt to reflect variable name, such as “qchanr” (river flow), or “qhocrn” (river discharge into the ocean).

For observational data, the hydrologist can get data from the Gravity Recovery and Climate Mission [15] and the Tropical Rainfall Measuring Mission [16] from the National Aeronautics and Space Administration (NASA) to validate the outputs of the climate model simulation. These datasets contain remote sensing imagery for tropical precipitation and storage. Ground stream flow data is available from the USGS. Fertilizer input and water-quality measurements may come from the EPA and the US Department of Agriculture.

Biogeochemical data is available to the hydrologist from the NASA-sponsored Distributed Active Archive Center at the Oak Ridge National Laboratory³(ORNL DAAC). This center holds about 1,000 datasets amounting to 2 Terabytes relevant to biogeochemical dynamics, ecological data, and environmental processes, as well as 60 TB of land product data subsets (measurements of surface radiance, reflectance, emissivity, and temperature) from the Moderate Resolution Imaging Spectroradiometer (MODIS) Instrument aboard the Terra and Aqua satellites.

A scientific user may typically be familiar with computational climate datasets, such as those found in ESGF, or with observational earth and ecological science datasets such as those found in the ORNL DAAC, but not both. Both types of sources currently present their data in faceted searches along attributes such as Project, Model, Experiment, Product, Variable Name, and Ensemble for ESG, and Parameter, Sensor, Topic, Project, Keywords in the ORNL DAAC. A faceted search exposes different views of a search result set along attributes contained in structured metadata. Note that in computational data the facet “Experiment” denotes experiments “in silico.” In the observational data, one also finds “Models,” a term typically reserved for simulations, where datasets are used in assessments and policy studies and simulate ecological systems: observational data can also be the result of simulations.

Thus, data solutions to the scientific question require the use of heterogeneous data. The hydrologist will need to search for datasets from different data centers to discover useful data. Each data domain has its own portal, its own metadata formats, and its own query-building methods for obtaining datasets. The exact definition of variables and observational parameters may require substantial searches for unfamiliar topics. In order to advance investigation of climate change, scientists need access to formal descriptions of the multiple objects present in each activity and to tools that permit seamless searches across all types of data. The next section presents tools that enable heterogeneous data access and improve searches.

5. Design and implementation

To enable data access and to facilitate searches, we integrated several existing technologies and developed an added module for query expansion using Earth Sciences domain ontologies. Our system is composed of the following components, which we describe in detail in the rest of this section:

- The National Center for Biomedical Ontology (NCBO) BioPortal ontology repository.
- The Mercury search engine

² <http://www.earthsystemgrid.org/>

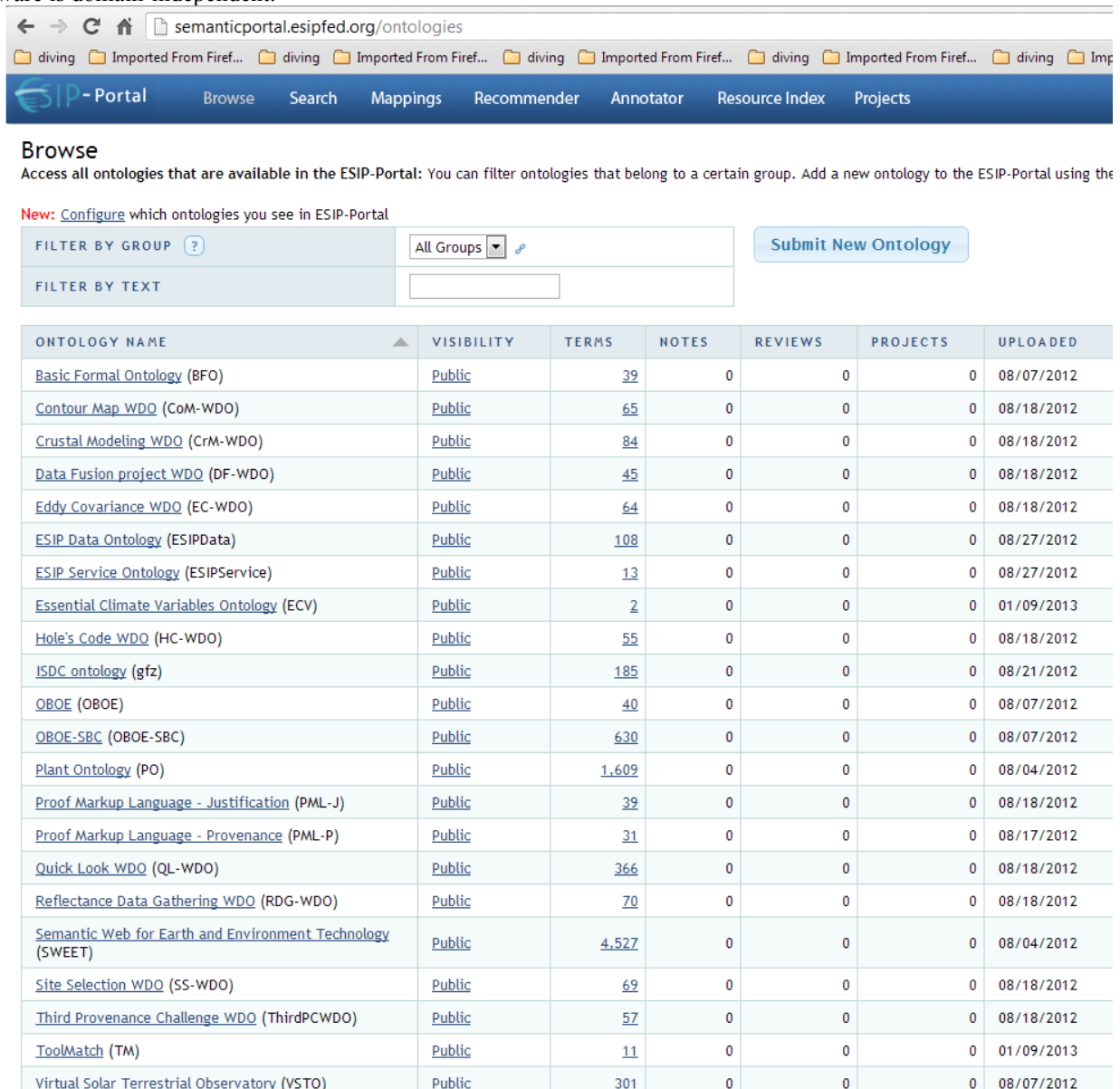
³ <http://daac.ornl.gov/>

- Our added module providing programmatic access from Mercury to ontology terms stored in our BioPortal instance.

A key factor in integrating these components was the existence of open source APIs.

5.1. The NCBO BioPortal and its Virtual Appliances

The NCBO BioPortal is a community-based ontology repository [17, 18]. BioPortal allows users to browse ontologies and to search for specific ontologies that have terms that are relevant for their work. The mappings between ontologies 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. While the instance of BioPortal that runs at NCBO is a repository of biomedical ontologies—with more than 300 of them at the time of this writing—the BioPortal software is domain-independent.



Browse

Access all ontologies that are available in the ESIP-Portal: You can filter ontologies that belong to a certain group. Add a new ontology to the ESIP-Portal using the

New: [Configure](#) which ontologies you see in ESIP-Portal

FILTER BY GROUP ? All Groups +

FILTER BY TEXT

[Submit New Ontology](#)

ONTOLOGY NAME	VISIBILITY	TERMS	NOTES	REVIEWS	PROJECTS	UPLOADED
Basic Formal Ontology (BFO)	Public	39	0	0	0	08/07/2012
Contour Map WDO (CoM-WDO)	Public	65	0	0	0	08/18/2012
Crustal Modeling WDO (CrM-WDO)	Public	84	0	0	0	08/18/2012
Data Fusion project WDO (DF-WDO)	Public	45	0	0	0	08/18/2012
Eddy Covariance WDO (EC-WDO)	Public	64	0	0	0	08/18/2012
ESIP Data Ontology (ESIPData)	Public	108	0	0	0	08/27/2012
ESIP Service Ontology (ESIPService)	Public	13	0	0	0	08/27/2012
Essential Climate Variables Ontology (ECV)	Public	2	0	0	0	01/09/2013
Hole's Code WDO (HC-WDO)	Public	55	0	0	0	08/18/2012
ISDC ontology (gfr)	Public	185	0	0	0	08/21/2012
OBOE (OBOE)	Public	40	0	0	0	08/07/2012
OBOE-SBC (OBOE-SBC)	Public	630	0	0	0	08/07/2012
Plant Ontology (PO)	Public	1,609	0	0	0	08/04/2012
Proof Markup Language - Justification (PML-J)	Public	39	0	0	0	08/18/2012
Proof Markup Language - Provenance (PML-P)	Public	31	0	0	0	08/17/2012
Quick Look WDO (QL-WDO)	Public	366	0	0	0	08/18/2012
Reflectance Data Gathering WDO (RDG-WDO)	Public	70	0	0	0	08/18/2012
Semantic Web for Earth and Environment Technology (SWEET)	Public	4,527	0	0	0	08/04/2012
Site Selection WDO (SS-WDO)	Public	69	0	0	0	08/18/2012
Third Provenance Challenge WDO (ThirdPCWDO)	Public	57	0	0	0	08/18/2012
ToolMatch (TM)	Public	11	0	0	0	01/09/2013
Virtual Solar Terrestrial Observatory (VSTO)	Public	301	0	0	0	08/07/2012

Figure 3 The ESIP Semantic Portal lists ontologies relevant to Earth sciences. It contains the SWEET ontology and many others. Users can browse each ontology hierarchy or search for terms of interest across all ontologies.

For the communities that want to run their own ontology repositories using the BioPortal code base, the NCBO team generates a *Virtual Appliance (VA)*—a packaged copy of the web-server software that other communities can install and maintain. These communities use the repository to share and access ontologies that are relevant to their domain. The Earth Science Information Partners Federation (ESIP) Semantic Portal⁴ deployed such a VA on the Amazon EC2 cloud node procured by ESIP [19]. Figure 3 shows ontologies uploaded in this repository. In addition to SWEET 2.3, the ESIP ontology collection includes the Plant Ontology, which describes structure and developmental stages of a plant [20], and the Extensible Observational Ontology (OBOE) for representing scientific observations and measurements [21], and its extension to represent ecological and environmental data.

The NCBO BioPortal at ORNL DAAC is another such installation behind a firewall. It contains the same version of the SWEET ontology as the ESIP ontology portal. We used the ORNL DAAC instance in our added module for ease of implementation and evaluation.

5.2. The Mercury tool: aggregating metadata

Mercury is a tool for distributed metadata harvesting, search, and retrieval originally developed for NASA. Mercury is currently used by projects funded by NASA, USGS, and U.S. Department of Energy (DOE) [22]. More than 30,000 scientists use Mercury each month. Mercury provides a single portal to search quickly for data and information contained in disparate data-management systems. It collects metadata and key data from contributing project servers distributed around the world and builds a centralized index. Mercury allows data providers to advertise the availability of their data and maintain complete control and ownership of that data. Figure 4 shows a diagram of the Mercury architecture including our added BioPortal module.

Mercury currently provides access to over 100,000 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, XML, Library of Congress protocols Z39.50⁵ and Search/Retrieve via URL⁶, and Amazon subsidiary A9's OpenSearch⁷.

The Mercury architecture 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 in a publicly accessible place such as a web directory or FTP directory. Mercury then harvests these metadata, builds the centralized index, and makes it 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) [23] and other web-based harvesting techniques.

The Mercury search interfaces allow users to perform simple, attribute-based, spatial and temporal searches across these metadata sources. The Mercury repository of metadata for distributed data sources provides low latency search results to the user. For instance a full-text search of 70,000 XML documents returned 48 records in 90 milliseconds; a fielded search of the same collection returned 7 documents in 122 milliseconds [22].

⁴ <http://semanticportal.esipfed.org/>

⁵ <http://old.cni.org/pub/NISO/docs/z39.50-brochure/>

⁶ <http://www.loc.gov/standards/sru/>

⁷ <http://www.opensearch.org/>

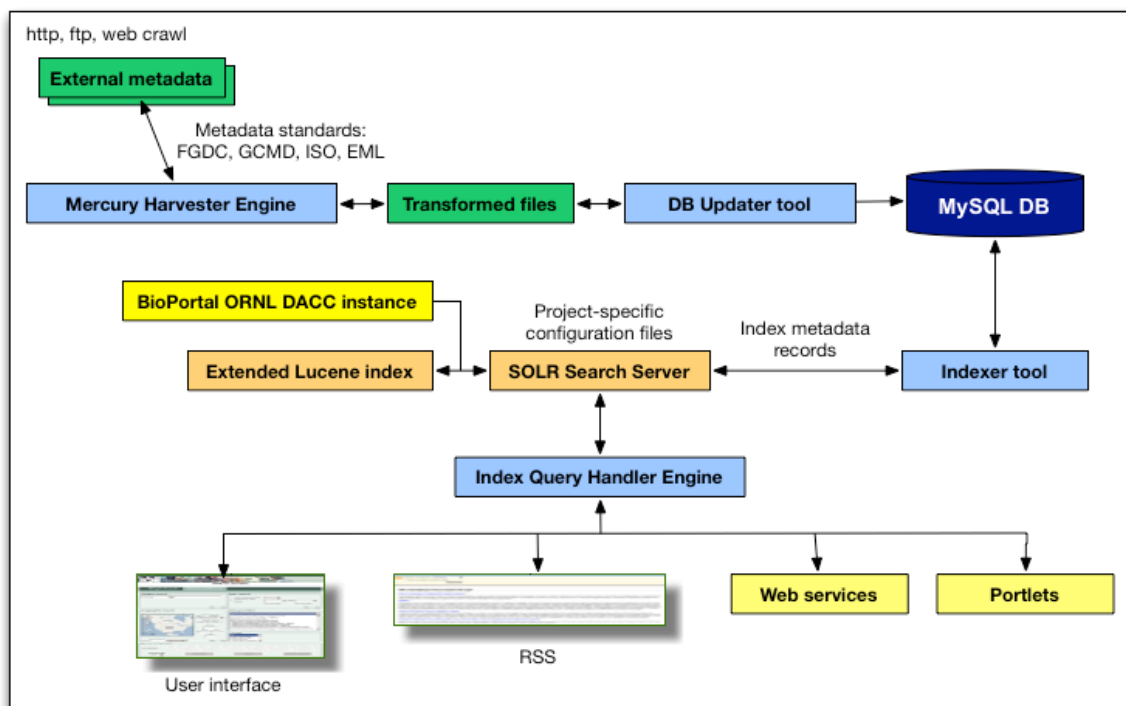


Figure 4: Architecture of the Mercury Search Engine and its integration with BioPortal ORNL DACC instance. Blue boxes indicate reusable software components. Green boxes are metadata files. Yellow boxes are external services. The Mercury Search service calls BioPortal REST services to add ontology knowledge to the queries

Mercury’s query engine is built using a service-oriented architecture, which includes a rich user interface. This interface allows users to perform various types of search capabilities, including 1) simple search, which performs a full text search, 2) advanced search, which allows users to search against controlled-vocabulary keywords, time period, spatial extent and data provider information, and 3) web browser tree search, which enables a drill-down through the metadata facets using a hierarchical keyword tree.

5.3. Adding semantics and integrating components

With the breadth of sciences represented within the Mercury metadata records, scientists can address some key interdisciplinary scientific challenges related to climate change and its environmental and ecological impacts, such as carbon sequestration and mitigation. However, the wealth of data and metadata also makes it difficult to pinpoint the datasets that are relevant to particular scientific inquiries.

We have applied semantic technologies—ontologies, in particular—to improve search results. There are several reasons for using this approach. First, simply using popularity determined by pointing links to provide a high ranking to a search result, as with Google, typically is not useful in the case of scientific data queries. Each scientific inquiry tends to be unique, and datasets are not directly indexed so that result ranking may not be useful. Thus, we must be able to improve search results based on the *meaning* of the data descriptions. Ontologies represent such *meaning* in a machine-readable way. Second, scientific queries are unlike everyday queries because they return specific datasets, which themselves have numerous parameters that may or may not be exposed to a

general search engine. For example, the Earth System Grid Federation (ESGF) gateway exposes 368 variables to search. Third, each domain science has its own terminology, more or less curated and consensual, and with various degrees of standardization. The same term may refer to different linguistic or scientific objects across domains (semantic plurality), and different terms mean the same thing (synonymy). For all these reasons, we decided to use scientific ontologies because they can provide a context for search results, in a way that keywords never will.

We used the SWEET ontologies to improve the results of the Mercury search interface. The ontologies provide context by linking individual keywords to a scientific realm and suggest additional keywords for searches. We designed an ontology service that allows integration of ontology terms into search results. The Mercury search system passes its search terms to the VA, which returns one or several matching terms through the REST interface. The user can choose any of these as additional search term for Mercury, or directly display the results indicated by the ontology sub-class terms. For example, an ontology-based search on biomass returns the keywords “biomass” and “litter” because litter is a sub-class of biomass in SWEET.

6. Results

The ontology service provides domain context, parameter attribute, and entity annotations to the Mercury search system. Mercury user interface uses a faceted search approach; we present the ontology results to the user in the same user interface (Figure 5). The search term used in this figure is “biomass.” The five top boxes (“Filter by”) show the faceted search results without semantic search. The bottom three boxes (“Ontology”) present the results of the semantic search.

Unlike a faceted search that highlights attributes within a set of results but cannot enlarge the set, the semantic solution can implement both restrictions and expansions of the initial set of results. In our semantic search, there are four new dimensions enabled by ontologies: Ontology Concepts, Ontology Super-classes, Ontology Sub-classes and Filter by keywords and all sub-classes. *Ontology concepts* present each search term within the ontological hierarchy. *Ontology Super-classes* shows the hierarchical level one level up and *Ontology sub-classes*—one level down. To display the facet “Filter by keywords and all sub-classes” the ontology service sends the sub-class terms to Solr, which returns links to datasets of interest (not shown in the figure).

6.1. Using ontologies to improve context

Recall the scenario that we described in Section 2. Our hydrologist will need to search for datasets annotated with “biomass” because she wants to analyze the transport of biochemical species in the river flow. She will search for datasets containing the term “biomass.” A Mercury search using controlled vocabulary keywords returns 35 datasets, a full-text search returns 187 datasets. A search for “biomass OR humus” (a type of biomass) returns 192 datasets, indicating that 5 potentially relevant datasets are not included in the search on biomass. Querying the SWEET ontologies through BioPortal’s REST API, the ontology service exposes “humus” as an additional search term for Mercury in the first discovery session about “biomass.” Humus is a sub-class of biomass in SWEET. Thus, the semantic search returns the five additional datasets without the user having to know about specific types of biomass. “Biomass” also acquires scientific context when the ontology service exposes that it can be a form of Energy Storage and a Living Entity. These examples demonstrate how ontologies help expand the search and provide scientific context for the search terms.

6.2. Using ontologies to reduce the number of search results

“Carbon” is another popular search term in Mercury, since the increase in the concentration of carbon dioxide in the atmosphere is considered a potential factor of climate change. A Mercury search for “carbon” returns 264 datasets from the ORNL DAAC. With the ontology service integrating the results of an ontology

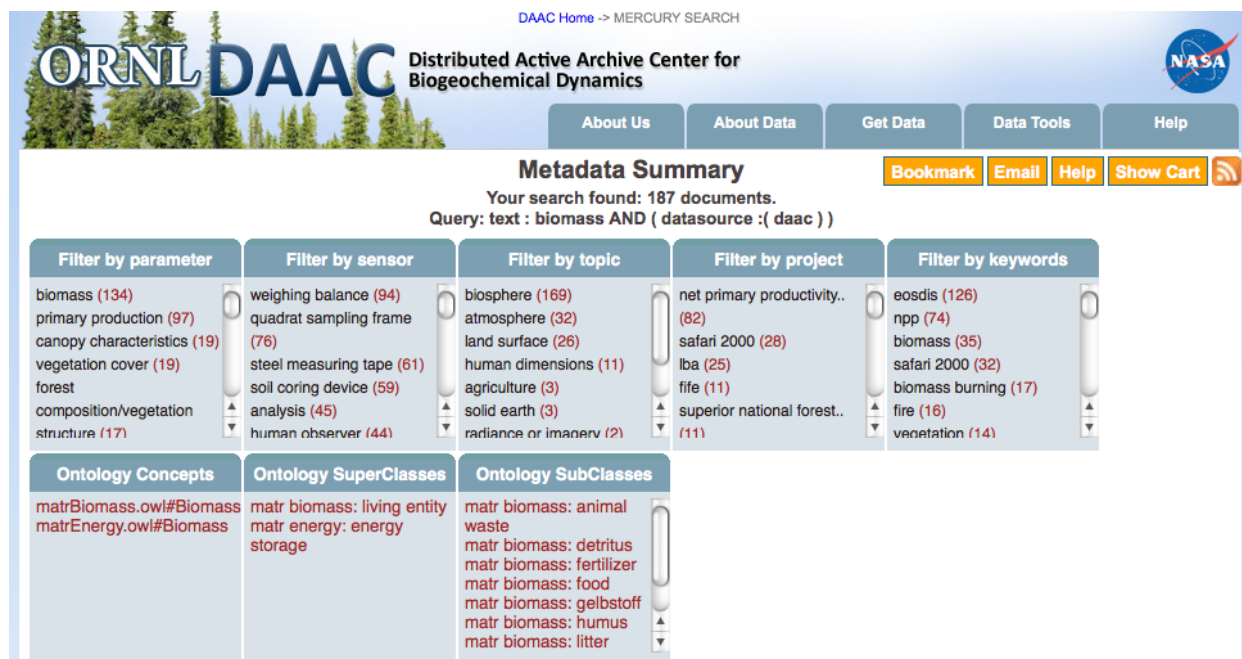


Figure 5: User interface for the semantic search in ORNL DACC. The user has searched for “biomass” and the interface suggest additional related terms based on the ontology search. The five tabs named “Filter” display Mercury search results. The three tabs named “Ontology” display the results of the ontology-based search obtained with our prototype

search into the faceted search, “carbon” acquires a scientific context and additional query terms that can be used to reduce the scope of the original search. For example, the individual in one of the ontologies, “stateTimeGeologic2:Carboniferous,” links results to datasets relevant to geological times (paleo-environmental science), while the sub-class “carbon offset” links to datasets relevant to “human environmental control” and “human activity.” In addition, “offset” is not a facet offered by the Mercury search system but the ontology search suggests this sub-class to reduce the result set further. Limiting the search to both “carbon” and “offset” produces only two results.

While ontologies provide additional search terms, we also use the ontology structure to enable the user to filter the results in a meaningful way.

6.3. Analyzing the coverage of ontologies

Using simple term matching, we evaluated how well the terms in the SWEET ontologies cover the top 100 controlled-vocabulary keywords that were used for indexing datasets in Mercury. “Biomass” is the top keyword currently indexing 138 datasets. Figure 6 shows the results of this evaluation. 21 of the top 100 keywords do not appear in the ontologies. Thus, 79% of the top 100 keywords in Mercury have at least one match in the selected ontologies. At the long tail of the distribution one keyword (water) has 38 matches, and two (air, carbon) have 28 matches. A fifth of the Mercury keywords do not appear in the SWEET ontologies. Thus, the scientific community needs to develop additional ontologies to enhance the keyword collection adequately.

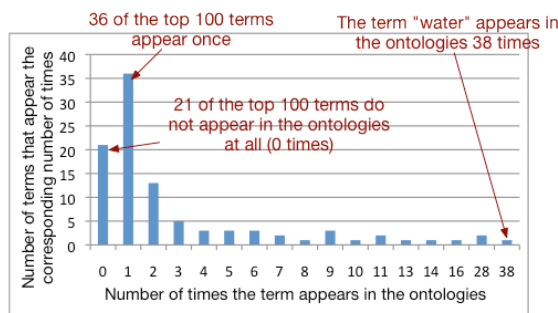


Figure 6 Ontology coverage of the top 100 controlled-vocabulary keywords.

7. Related work

Researchers in the Semantic Web community have studied *semantic search* and a variety of approaches to it. A recent survey [24] presented a general model for semantic search and identified different types of semantic search. In general, there are two key approaches. In one, the (linked) data is represented in RDF or OWL and the search engine provides access to a collection of such data, either through keyword search or through SPARQL (e.g., SWSE [25], or Sindice [26]). Uren and colleagues provide a survey of this type of semantic-search engines [27].

The second class of semantic-search applications are document-retrieval applications that use semantics to expand or constrain the user query (e.g., [28, 29]). The query expansion method [30] uses ontology terms related to terms of the original query as additional search terms. Related ontology terms can specify synonyms, sub- or super-classes for query terms, thus providing additional search terms that are not linguistically related to query terms. Semantic technologies, such as ontologies, help improve search results by adding these additional search terms and thus potentially increasing the number of returned results.

The application that we describe here is closer to the second category as we use ontologies for query expansion using super-classes and sub-classes. It provides access to heterogeneous collections of structured data, but this data is not represented in Semantic Web formats, thus does not fall into the first category. At the same time, it uses semantics on the “front-end,” augmenting the user query, but we use this query expansion to access structured data and not a set of documents. Thus, to the best of our knowledge, the application that we describe is unlike many semantic-search applications because it uses semantics on the query side, provides access to structured data, but not in RDF and OWL format.

Kauppinen and colleagues frame the challenges of Linked Science in the form of an “executable paper” [2], with publication of validated and well-sourced data as one of the key requirements. Contributions to the First and Second Linked Science workshops⁸ [4] investigated several issues related to Linked Science and Linked Data but did not focus on semantic searches for structured datasets in dedicated archives. Researchers discussed: the requirements for Linked Science in the geo-physical sciences [31]; the use of rules for interactively mapping data sources in databases to ontology and generating RDF triples [32]; the need for trust in the data sources with an emphasis on formally describing the relationship between data and sources in bibliographic resources [33]; and challenges in the bioinformatics [34] and astronomy domains. Thus, researchers are actively addressing the trends

⁸ <http://linkedscience.org/events/lisc2011/>
<http://linkedscience.org/events/lisc2012/>

in Linked Science. Our effort describes our semantic work in the context of Linked Science and is complementary to the approaches described in these papers.

8. Discussion

Our approach to the investigation of climate change has led to the integration of search capabilities and the development of a semantic service for discovering multi-disciplinary datasets in Earth and Environmental sciences. Scientists can use our semantic service to discover new datasets that were not included in the original search results, thus expanding the original queries.

We used a BioPortal instance as a source for ontologies rather than a triple-store or an OWL API to process the ontologies for several reasons. First, the REST service interface that BioPortal provided was easy to integrate into the Mercury architecture. Second, ontology authors sometimes use idiosyncratic approaches to representing some features of their ontologies, such as preferred names or synonyms for terms. These lexical features are key to user searches but ontologies use different properties to represent them. BioPortal uses ontology metadata to extract these properties and provides its users with a single service call to access this information across all ontologies in a repository. Finally, BioPortal enables scientists to submit new ontologies through its web interface and these ontologies become available to the semantic search in Mercury. Thus, if a scientist discovers a new ontology that covers her domain of interest, she can add it to her set of ontologies to expand the meaningful results from her semantic search.

We set up the ESIP instance of BioPortal because this user community needs a stable ontology repository that covers the Earth and Environmental Sciences domains. This instance of BioPortal is accessible to users with all the functionality provided by BioPortal, including annotations, ontology extensions, and term mappings. New community additions to the ontologies made through this instance are directly accessible to the semantic service. ESIP is currently discussing curation mechanisms for ontologies in the ESIP portal.

However, our approach has several limitations. First, the faceted display becomes crowded very quickly and a more dynamic presentation of search results may be beneficial. Another, more serious, limitation is that the quality of the newly discovered metadata is contingent on the quality of the ontologies used in our implementation. BioPortal curates the ontologies by enforcing compliance with ontology language standards and resolving relationships and axioms to detect potential conflicts, but it cannot check for coverage or correctness in terms of domain expertise. Search terms and thesaurus keywords in Mercury may be absent from current ontologies, or the ontology classification may not bring additional information that is not already presented by the faceted terms. However, as semantic technologies mature, more substantial ontologies become available in many scientific domains.

9. Conclusion and Future Directions

We have made several initial steps in order to address the limitation on coverage and quality of the ontologies. We will use the features that are currently available in BioPortal to solicit feedback and to provide additional information about the ontologies. Specifically, BioPortal includes a comment field for each ontology term that users can edit. ESIP members can take advantage of this feature to resolve conflicts and to propose new terms. Second, the ESIP Semantic Web portal team is currently working on manual evaluation of the coverage of ontologies. The team plans to submit proposals and annotations of terms to the ESIP community for approval. Finally we plan to use the mapping function of BioPortal for creating mappings between terms in the ontologies, thus helping to extend coverage.

The solution that we presented in this paper leverages the federated search capabilities in Mercury that collect metadata records from several scientific domains, and the storage, access and curation functionality of BioPortal. Our solution provides guidance on how to leverage semantic capabilities for improving search results.

Use of ontologies—even lightweight ones—provides a path for helping domain experts find the information that they need from heterogeneous datasets. We demonstrated that the tools that are already available today enabled us, with minimal additional effort, to build on two mature systems and to find relevant datasets for interdisciplinary inquiries. The paper thus indicates a direction for linking environmental, ecological and biological sciences.

Acknowledgments

This work has been in part performed at Oak Ridge National Laboratory, Managed by UT Battelle, LLC under Contract No. De-AC05-00OR22725 for the U.S. Department of Energy

References

1. Raskin, R. *Guide to SWEET ontologies*. NASA/Jet Propulsion Lab, Pasadena, CA, USA, Available at: <http://sweet.jpl.nasa.gov/guide.doc> (last accessed: May 2011) 2006.
2. Kauppinen, T. and G. de Espindola. *Linked open science? Communicating, sharing and evaluating data, methods and results for executable papers*. in *International Conference on Computational Science, ICCS 2011*. 2011.
3. Kauppinen, T., L. Pouchard, and C. Kessler. *Proceedings of the First International Linked Science Workshop*. in *Second International Linked Science Workshop*. 2012. Boston, MA: CEUR.
4. Kauppinen, T., L.C. Pouchard, and C. Kessler, eds. *Proceedings of the First International Workshop on Linked Science (LISC 2011)*. Vol. CEUR Workshop Proceedings 783. 2011.
5. Michener, W.K., et al., *Participatory design of DataONE—enabling cyberinfrastructure for the biological and environmental sciences*. *Ecological Informatics*, 2012. **11**: p. 5-15.
6. Fink, D., et al., *Spatiotemporal exploratory models for broad-scale survey data*. *Ecological Applications*, 2010. **20**(8): p. 2131-2147.
7. Kelling, S., et al., *Estimating species distributions—across space, through time, and with features of the environment*, in *The DATA Bonanza: Improving Knowledge Discovery in Science, Engineering, and Business*, M. Atkinson, et al., Editors. 2011, John Wiley & Sons, Inc.: Hoboken, NJ. p. 441-458.
8. Gil, Y., V. Ratankar, and P. Hanson, *Organic data publishing: A novel approach to scientific data sharing*, in *2nd International Workshop on Linked Science (LISC 2012)*, T. Kauppinen, C. Kessler, and L. Pouchard, Editors. 2012, CEUR: Boston, MA.
9. Patton, E.W., et al., *Assessing health effects of water pollution using a semantic water quality portal*, in *10th International Semantic Web Conference* L. Aroyo, et al., Editors. 2011, Springer: Bonn, Germany.
10. Raskin, R.G. and M.J. Pan, *Knowledge representation in the semantic web for Earth and environmental terminology (SWEET)*. *Computers & Geosciences*, 2005. **31**(9): p. 1119-1125.
11. Buttigieg, P.L., et al., *The environment ontology: contextualising biological and biomedical entities*. *Journal of biomedical semantics*, 2013. **4**(1): p. 43.
12. Gent, P.R., et al., *The Community Climate System Model Version 4*. *Journal of Climate*, 2011. **24**(19): p. 4973-4991.
13. Doney, S.C., *The Growing human footprint on coastal and open-ocean biogeochemistry*. *Science*, 2010. **328**(5985): p. 1512-1516.
14. Bernholdt, D., et al., *The Earth System Grid: Supporting the next generation of climate modeling research*. *Proceedings of the IEEE*, 2005. **93**(3): p. 485-495.
15. Tapley, B.D., et al., *The gravity recovery and climate experiment: Mission overview and early results*. *Geophysical Research Letters*, 2004. **31**(9).
16. Theon, J.S., *The Tropical Rainfall Measuring Mission (Trmm)*. *Remote Sensing of Earths Surface and Atmosphere*, 1993. **14**(3): p. 159-165.

17. Musen, M.A., et al., *The National Center for Biomedical Ontology*. Journal of American Medical Informatics Association (JAMIA), 2011.
18. Whetzel, P.L., et al., *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.
19. Pouchard, L., M. Huhns, and A. Depriest, *Lessons learned in deploying a cloud-based knowledge platform for the ESIP federation*, in *American Geo-physical Union Fall meeting*. 2012: San Francisco.
20. Bruskiewich, R., et al., *The Plant Ontology (TM) Consortium and plant ontologies*. Comparative and Functional Genomics, 2002. **3**(2): p. 137-142.
21. Madin, J., et al., *An ontology for describing and synthesizing ecological observation data*. Ecological Informatics, 2007. **2**(3): p. 279-296.
22. Devarakonda, R., et al., *Mercury: reusable metadata management, data discovery, and access system*. Earth Science Informatics, 2010. **3**(1-2): p. 87-94.
23. Devarakonda, R., et al., *Data sharing and retrieval using OAI-PMH*. Earth Science Informatics, 2011. **4**(1): p. 1-5.
24. Tran, T., D.M. Herzig, and G. Ladwig, *SemSearchPro--Using semantics throughout the search process*. Web Semantics: Science, Services and Agents on the World Wide Web, 2011. **9**(4): p. 349-364.
25. Hogan, A., et al., *Searching and browsing Linked Data with SWSE: The Semantic Web Search Engine*. Web Semantics: Science, Services and Agents on the World Wide Web, 2011. **9**(4): p. 365-401.
26. Tummarello, G., R. Delbru, and E. Oren. *Sindice.com: Weaving the Open Linked Data*. in *International Semantic Web Conference (ISWC-2007)*. 2007. Busan, Korea: Springer Berlin / Heidelberg.
27. Uren, V., et al., *The usability of semantic search tools: a review*. The Knowledge Engineering Review, 2007. **22**(04): p. 361-377.
28. Castells, P., M. Fernandez, and D. Vallet, *An adaptation of the vector-space model for ontology-based information retrieval*. IEEE Transactions on Knowledge and Data Engineering, 2007. **19**(2): p. 261-272.
29. Chu-Carroll, J., et al. *Semantic search via XML fragments: a high-precision approach to IR*. 2006. ACM.
30. Navigli, R. and P. Velardi. *An analysis of ontology-based query expansion strategies*. in *Proceedings of the 14th European Conference on Machine Learning, Workshop on Adaptive Text Extraction and Mining, Cavtat-Dubrovnik, Croatia*. 2003.
31. Mäs, S., et al. *Linking the Outcomes of Scientific Research: Requirements from the Perspective of Geosciences*. in *First International Workshop on Linked Science (LISC2011) at ISWC 2011*. 2011. Bonn, Germany.
32. Knoblock, C.A., et al. *Interactively Mapping Data Sources into the Semantic Web*. in *First International Workshop on Linked Science (LISC2011) at ISWC 2011*. 2011. Bonn, Germany.
33. McCusker, J.P., et al. *Where did you hear that? Information and the Sources They Come From*. in *First International Workshop on Linked Science (LISC2011) at ISWC 2011*. 2011. Bonn, Germany.
34. Vision, T., et al. *Similarity Between Semantic Description Sets: Addressing Needs Beyond Data Integration*. in *First International Workshop on Linked Science (LISC2011) at ISWC 2011*. 2011. Bonn, Germany.

