Earth Cube Technical Solution Paper - Semantic Framework for the EarthCube CI

Nancy Wiegand, Hank Revercomb (Director)
Space Science and Engineering Center
University of Wisconsin-Madison

Introduction

The EarthCube vision is for transformative science to be enabled by a cyberinfrastructure that makes available and integrates data and methods across earth science domains. This is a challenging goal but a promising approach to solve complex earth science problems. This paper discusses the need for semantic technologies in the EarthCube cyberinfrastructure to help achieve this goal. Areas needing formal semantic knowledge are outlined.

The EarthCube cyberinfrastructure will be a distributed system accessible over the Web that facilitates complex problem solving through the availability and integration of diverse but related data, along with available processes to manipulate and analyze the data. Data and processes would mostly stay local and be accessible through web services. Examples from UW-Madison's Space Science and Engineering Center (SSEC) are data being made more accessible using the Open Geospatial Consortium's (www.opengeospatial.org) web service capabilities and the VisAD/IDV/McIDAS V software that is Web-enabled.

The data and processes in EarthCube will span earth science domains, each of which has its own concepts, terminology, and methods. To enable analysis and integration across such diverse data and processes, additional background knowledge frameworks are needed. Differences in terms and concepts need to be resolved both within a domain and across domains. That is, semantic heterogeneity occurs even within disciplines and here needs to be resolved across the various earth science domains if data sets will be combined in some manner or be part of the same process. It is very typical that alternate terms are used for the same concept by different data providers or programmers. Background knowledge bases would relate terms in a specified manner, such as by subclassing, by type, or by designating any specific relationship (e.g., part of or observed by). Even basic unifying concepts, such as location and time, have different representations that may need to be resolved in concept.

Further, in addition to background knowledge bases that resolve individual data source terms or parameter identifiers, **a top level knowledge framework covering all EarthCube data and processes would be valuable**. For example, to facilitate access to the CI, a background content design can guide the user to data and processes. Such a high level framework needs to be more than just a complete categorization of the types of data and processes from the different domains, however. Instead, such a knowledge base of background information would contain, for example, relationships between types of data, types of operations that could be applied to the data, what types of tasks use this data, and so forth. Recording this information using formal semantic technologies will help in organization and discovery. Further, formalization of this information will allow automated access and processing.

Other formalizations of background knowledge could include those for processes or tasks. The cyberinfrastructure will likely include capabilities to perform specific prior-identified complex tasks that are known to be useful. Such tasks need coordination between concepts. Conceptual models could be designed that contain the types of data needed for important or typical tasks, as well as the data manipulations involved in the task. Such models could be used for automation in performing the task.

Further, the CI must be designed to enable solving grand challenges, for which solutions involving data and processes are not yet known. Enough background concept information should be available to allow ad hoc operations for new combinations of information. Developing knowledge bases using formal semantics will facilitate these needs. That is why we propose including formal semantic background knowledge at the top level of design for the entire cyberinfrastructure for EarthCube as well as at lower levels, such as for individual data sources. Again, this background knowledge will help the user in the discovery of data and also help in using data in analyses and processes. The more background knowledge designed for the CI, the more automation of search or workflows is possible.

Formal Ontologies

Although background knowledge could recorded in tables or in other ways, formal semantic technologies are now being standardized by the World Wide Web Consortium (W3C, w3c.org) and are gaining interest and being applied. Standard languages and tools are being developed. For example, the W3C has released standards for expressing and querying ontologies. Ontologies are formal representations of terms, concepts, and relationships that occur in a domain. OWL is the W3C standard for a formal representation of an ontology over which reasoning can be done. An OWL ontology is an RDF graph. RDF is a standard for a triple format to represent information in a simple manner. In RDF, each component of the triple is identified by a unique Web identifier (URI), which enables linkages across different knowledge bases that use the same URI (or that have URIs declared as equivalent). SPARQL is the W3C standard query language for RDF. SPARQL allows querying to be done directly over the Web. GeoSPARQL is currently under review as a spatially-enabled Web RDF query language with spatial functions.

Various ontologies in earth science domains already exist, but more are needed. Ontology work in the geosciences has been done as part of the GEON project (www.**geon**grid.org/) and also reported in (Sinha 2006). Work on ontologies for Virtual Observatories has been done (e.g., Fox et al., 2009). Also, the NASA SWEET (Semantic Web for Earth and Environmental Terminology, http://sweet.jpl.nasa.gov/ontology/) ontologies can serve as a starting point to represent concepts in each of the earth science domains and to form overarching ontologies to combine concepts across the earth sciences (Figure 1).

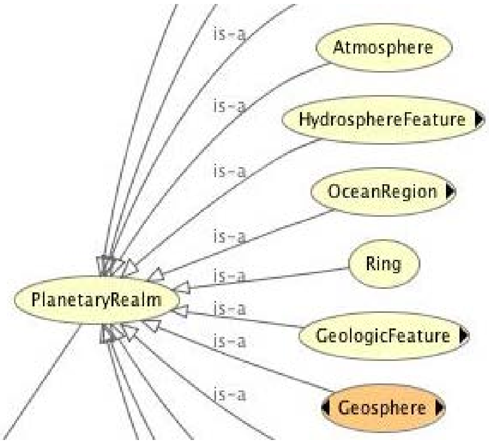

Figure 1. Part of NASA's SWEET Ontology for realmGeolBasin.owl

Topics

As discussed, formally represented background knowledge will be used in many aspects of the design for the EarthCube CI. In addition to a comprehensive background ontology for the CI itself, ontologies are needed in the design for metadata, searching for data, processes on data, and querying data. Another issue is the placement and access of ontologies in the overall CI architecture, including ontology repositories. Finally, as part of the Semantic Web vision, earth science data could be put into RDF to be part of linked open data. These topics are discussed as follows.

Metadata. Metadata is used to guide discovery and use of data. Geospatial data, which is related to earth science data, have FGDC and ISO 19115 metadata standards to describe the overt characteristics of the data set. This information is used in catalogs/portals to help search for data, as in Geospatial One-Stop (geodata.gov, now being migrated to data.gov). However, such metadata was not specifically created for web searching, and, as a result, hundreds of data sets are returned given a <location, theme_keyword> search pair. For an EarthCube CI, metadata elements need to be designed to differentiate specific data and services available. Creating targeted metadata could be done similarly to creating collaborative ontologies by having initial in-person facilitated sessions to delineate types of data, types of processes on the data, types of tasks using the data, other uses of data, types of processes in general, and so on. Virtual sessions can then fine-tune the result. It is important that metadata be designed to be extensible, however, to be able to accommodate new types of data, processes, and uses not initially planned.

Search. Search in a cyberinfrastructure typically is done over metadata (versus querying, which is typically done over the actual data sources). With appropriately designed metadata elements, data and processes should be able to be more precisely found in the CI.

In addition, background ontologies are needed to expand search and for browsing. Ontologies for metadata element values used by metadata creators can be used for term expansion. Such expansion is now typically limited in search systems to synonym lookups. However, the vision here is to have thorough and inter-related ontologies of domain and cross domain information to be able to more precisely pinpoint what the user is seeking.

Ontology Repositories. Ontology repositories are now being created by the Semantic Web community (e.g., the BioPortal (http://bioportal.bioontology.org/) and the Open Ontology Repository (OOR, http://openontologyrepository.org/)). The OOR is starting to be used to collect geospatial ontologies as part of an NSF INTEROP project (www.socop.org). Ontology repositories have capabilities to store ontologies and mappings between ontologies, and they implement browsing and search for ontologies. Ontologies for EarthCube could be stored in existing repositories, which will be distributed, or a dedicated EarthCube ontology repository. In either case, the cyberinfrastructure needs to be able to access ontologies and know when and how to use them.

Query and Processing. Querying and processing of data are the most complex data functions in the CI. Because standards are not often available or used, querying across data sets even in the same domain requires mappings or resolutions of terms. Similarly, parameters need to match across software that processes data. For an example, as an abstraction over different data formats, the Space Science and Engineering's VisAD software developers created an extensible grammar or data model to allow many different file formats as well as input parameters. However, in general, if parameter terms vary across data that need to be combined, then mappings will be needed.

To solve semantic heterogeneity for querying and processing, many ontologies are likely necessary. Some of these can be dedicated to particular domains or processes. Further, an upper level, overarching ontology that combines various domain ontologies will enable automatic correlation between the different earth science domains.

Creating ontologies needed for querying and processing is related to the discussion above for creating metadata elements. However, metadata elements targeted for searching likely will not go into the detail needed for resolving individual data source terms for querying. Again, in-person and remote facilitation would be done across scientists in the earth science domains to understand terms and concepts and create mappings and linkages.

As another design consideration, if data sources themselves, attributes, parameters, metadata files, and metadata elements are explicitly linked to ontologies, then more automated query expansion or semantic resolution can be done. Explicitly linking such components to ontologies is currently not typically being done in systems, but we suggest that a CI does this. Otherwise, a user needs to know which ontologies are relevant for each particular need.

Architecture. Ontologies or other kinds of knowledge bases need to be fit into the CI architecture. The CI architecture is distributed with communication done by web services. There is access to remote or local data, metadata, ontology repositories, service repositories, etc. Code base modules will already exist for searching, querying, and running processes, for example. To add semantic capabilities to the architecture increases the complexity. There will be local and remote ontologies that need to be accessed for a variety of tasks as well as for resolving terms between individual data sets or similar processes. As a result, additional storage, access, and processing management are needed to access and use ontology information. This may be facilitated by the high level extensible background knowledge base describing the system as a whole as mentioned in the Introduction. Architectures including semantic elements are starting to appear, such as ICAN, iPlant, and INSPIRE, but more work is needed. Semantic architectures are currently being studied by the Open Geospatial Consortium (http://www.ogcnetwork.net/ows-8).

Linked Open Data. There is an initiative in the Semantic Web community to convert data to RDF and publish it in the linked data cloud (http://linkeddata.org/) (Figure 2). Many government documents, for example, are being published in RDF to be accessible over the Web as a method to open up the data to others as well as link it to other data, allowing more and new relationships to be discovered through the links. An example is the work being done for the contents of data.gov. As another example, data for The National Map (USGS) are currently being converted to RDF with a SPARQL endpoint set up (http://cegis.usgs.gov/projects.html). As part of the EarthCube CI, some of the earth science data can experimentally be put into RDF and either linked to the existing cloud or form its own cloud. This active and emerging linked open data initiative is only just starting but has potential for making data available in additional ways and for discovering new information by traversing links.

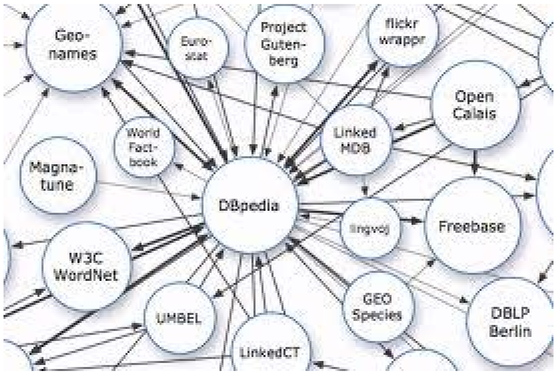

Figure 2. Part of the Linked Open Data Cloud

Summary

This paper delineates aspects of semantic work needed for an EarthCube CI. Because of the diverse and vast amounts of data, processes, and types of information expected to be available in the CI, formal knowledge bases describing and relating various aspects, from individual parameters and database attributes to a CI-level framework of components, will help in organization, searching, and ultimately, automation in the use of the cyberinfrastructure.

About SSEC

SSEC is a research and development center with primary focus on geophysical research and technology to enhance understanding of the atmosphere of Earth, the other planets in our Solar System, and the cosmos. SSEC researchers sometimes explore the universe from space and terrestrial-based telescopes, and probe other planets in our solar system, but more often they examine the Earth to gain information and insight into weather, climate, and other aspects of Earth's global environment. They develop new observing tools for spacecraft, aircraft, and ground-based platforms, and model atmospheric phenomena. They receive, manage and distribute huge amounts of geophysical data and develop software to visualize and manipulate these data for use by researchers and operational meteorologists all over the world.

Three related but conceptually distinct aspects of the above activities, refined during almost four decades serving the Earth Sciences community, form the basis for SSEC's Technology Solution papers:

1) Data abstraction, exemplified by the VisAD data model, as critical underpinning for interoperable processing, visualization and data exchange (this paper);
2) Broad community support through the Open Geospatial Consortium protocols; and
3) Semantic implications for EarthCube.

References

Fox, P., McGuinness, D.L., Cinquini, L., West, P., Garcia, J., Benedict, J., and Middleton, D. 2009. Ontology-supported Scientific Data Frameworks: The Virtual Solar-Terrestrial Observatory Experience. *Computers & Geosciences*, pp. 724-738.

Sinha, A.K. (editor). 2006. Geoinformatics: Data to Knowledge. The Geological Society of America, Inc., Special Paper 397. Boulder, CO