Enabling Semantic Interoperability for Earth Science Data

Rob Raskin, Michael Pan, and Chris Mattmann
Jet Propulsion Laboratory
California Institute of Technology
Pasadena, CA 91109

Abstract-**Data interoperability across heterogeneous systems can be hampered by differences in terminology, particularly when multiple scientific communities are involved. To reconcile differences in semantics, a common semantic framework was created through the development of Earth science ontologies. Such a shared understanding of concepts enables ontology-aware software tools to understand the meaning of terms in documents and web pages.**

This presentation updates last year's presentation on the Semantic Web for Earth and Environmental Terminology (SWEET). For the recent work, we incorporated concepts of other funded initiatives such as ESML, ESMF, grid computing, and OGC. We also created a system to update its knowledge base as needed, from gazetteers and other on-line Web sources. An accompanying search tool supports system-wide search and ultimately, a wide range of semantically-based web services.

I. Introduction

Earth system science data originate from many disciplines, spanning several community standards, terminologies, and data formats. Several initiatives are underway to develop a common infrastructure to improve data interoperability across the disciplines. Examples include the: Earth Science Markup Language (ESML), Earth Science Modeling Framework (ESMF), and the Open GIS consortium (OGC). Key to the success of these initiatives is the development of a common semantic framework. Such a framework enables dataset and science concepts to be understood by software tools. The framework goes beyond *data* interoperability by supporting the exchange of *conceptual knowledge* within and across these disciplines.

This framework can be achieved through the "Semantic Web" (Fensel, et al., 2003), an ambitious extension to the existing WWW environment, coordinated by the World Wide Consortium (W3C). The Semantic Web encodes common sense knowledge directly into web pages themselves, using broadly agreed upon namespaces and ontologies to define terms and their mutual relationships.

The motivation of our task is to improve semantic understanding of web resources by software tools, with specific application to discovery and use of Earth science data. Semantic understanding of text by automated tools is enabled through the combined use of *i)* ontologies and *ii)* software tools that can interpret the ontologies. An ontology is a formal representation of technical concepts and their interrelations in a form that supports domain knowledge. Generally, an ontology is hierarchical, with child concepts having explicit properties to specialize their parent concept(s).

A Semantic Web emerges if terms on web pages are associated with corresponding elements in ontologies. This is accomplished by placing an XML tag around a term to identify its associated ontology namespace. A search tool potentially can use these metadata tags to distinguish different uses of the same term (e.g. “fall” as a season vs. “fall” as a downward motion) to eliminate false hits. It also can locate resources without having an exact keyword match, because terms such as “El Nino” have an equivalent definition in terms of its defining scientific components.

To support potential Semantic Web activities, we developed a collection of ontologies for the Earth and environmental sciences and supporting areas. We created a common sense knowledge base of the Earth sciences using the Ontology Web Language (OWL) [1], a standard adopted by the W3C. We use these ontologies in a prototype search tool that improves performance by creating additional relevant search terms based on the underlying semantics. We demonstrate how such a knowledge base can be “virtual” by adding a wrapper around remote, dynamic data repositories.

II. Semantic interoperability

In the early days of computing, an initial level of data interoperability resulted when data structures (arrays) created on one computer system were *readable* by another computer. Data formats such as HDF emerged to extend this level of interoperability to more complex data structures & vendor platforms and enabled the preservation of variable *names*. The Internet later brought on protocols such as DODS [2], which supported modification of the data structure (subset extraction) during the transfer. Exchanges of this type say nothing about the scientific

interpretation of the data on the receiving end. A variable name is assigned to a data structure, but human intervention is required to make sense of it.

The HDF-EOS format remedied the semantic interoperability problem for *independent* variables by standardizing the naming convention of spatial and temporal parameters. The Open GIS Consortium (OGC) [3] provides a similar level of spatial/temporal interoperability problem in its Web Mapping Service (WMS) and Web Coverage Service (WCS) protocols. The HDF-EOS and OGC solutions enable a data seeker to query and access data by spatial/temporal parameters rather than by array row/columns (which would require human intervention). Thus a software tool understanding these conventions can access *any* HDF-EOS or OGC-compliant dataset and be guaranteed that the spatial-temporal interpretation is known.

Semantic interoperability for dependent variables has generally meant the use of controlled keywords. For instance, the NASA GCMD defines approximately 1000 controlled keywords, each with a dictionary definition. Such a representation does not support computer reasoning that would be required to respond to general queries or chain services together. It does not provide a rich expression of the relationship between the keywords and is not directly extendable by the user. This project addresses a more scalable solution to semantic interoperability in the context of the Earth sciences.

III. Ontology development

An ontology is a formal representation of technical concepts and their interrelations in a form that captures domain knowledge. Generally, an ontology is hierarchical, with child concepts having explicit properties to specialize their parent concept(s). Thus, "hydrosphere" is the parent concept of "surface water", which is a parent of "river", which is a parent of "Mississippi River", etc. In this paper, we describe our experiences with the development of Earth and environmental science ontologies.

In the initial year of ESTO funding, we created the Semantic Web for Earth and Environmental Terminology (SWEET) [4] to prototype how a Semantic Web can be implemented in the Earth sciences. We used the ~ 1000 terms in the Global Change Master Directory (GCMD) [5] as a starting point in manually populating the ontologies, but reorganized and expanded the concepts to form a scalable framework. Later, we incorporated an analogous keyword list of ~ 350 terms used in the Earth Science Modeling Framework (ESMF), based on a Standard name convention [6]. Our approach is to develop *faceted* ontologies, in which concepts are decomposed into their most basic parts, and combinable upon demand. Additional terms were added from other sources.

Earth Realm

The "spheres" of the Earth constitute an *EarthRealm* ontology, based upon the physical properties of the planet. Elements of this ontology include "atmosphere", "ocean", and "solid earth", and associated subrealms (such as "ocean floor" and "atmospheric boundary layer"). The subrealms generally are distinguished from their parent classes, based on the property of altitude, e.g., "troposphere" is the subclass of "atmosphere" where elevation is between 0 and 15 km.

Non-Living Element

This ontology includes the non-living building blocks on nature, such as: particles, electromagnetic radiation, and chemical compounds.

Living Element

This ontology includes plant and animal species. It was imported from the "biosphere" taxonomy of GCMD.

Physical Property

A separate ontology was developed for physical properties that might be associated with any component of *EarthRealm, NonLivingElements, or LivingElements*. *PhysicalProperties* include "temperature", "pressure", "height", "albedo", etc.

Units

Units are defined using Unidata's UDUnits. The resulting ontology includes conversion factors between various units. Prefixed units such as km are defined as a special case of m with appropriate conversion factor.

Numerical Entity

Numerical extents include: interval, point, 0, $\mathbf{R}^2$, … Numerical relations include: greaterThan, max, … We defined multidimensional concepts such as coordinate systems, mathematical operators, and functions.

Temporal Entity

Time is essentially a numerical scale with terminology specific to the temporal domain. We developed a time ontology in which temporal extents and relations are special cases of their numeric analogs. Temporal extents include: duration, season, century, 1996, … Temporal relations include: after, before, …

Spatial Entity

Space is essentially a 3-D numerical scale with terminology specific to the spatial domain. We developed a space ontology in which the spatial extents and relations are special cases of numeric extents and relations,

respectively. Spatial extents include: country, Antarctica, equator, … Spatial relations include: above, northOf, …

Phenomena

A phenomena ontology is used to define transient events. A phenomenon crosses bounds of other ontology elements. Examples include: hurricane, earthquake, El Nino, volcano, terrorist event, and each has associated *Time*, *Space*, *EarthRealm*s, *NonLivingElements*, *LivingElements*, etc. We also include specific instances of recent phenomena.

Human Activities

This ontology is included for representing impacts of environmental phenomena (commerce, fisheries, etc.)

Data

The data ontology provides support for dataset concepts, including representation, storage, modeling, format, resources, grid computing, and distribution. This ontology provides the namespace for semantic tags that may be included in an Earth Science Markup Language (ESML) [7] descriptor file, as described in the next section.

IV. ESML

The Earth Science Markup Language (ESML) combines an XML-based language for describing datasets with an API read library. Its XML tags are of two types: syntactic (for reading data) and semantic (for interpreting data). SWEET tags may be used to provide the semantic content of any ESML file.

V. ONTOLOGY LANGUAGES

An ontology is expressed using a *language* that is typically a specialization of XML. XML is widely supported by existing software tools and is platform-independent. The World Wide Consortium (W3C) has adopted two XML languages as its standard method of representing ontologies: the Resource Description Framework (RDF) and the Ontology Web Language (OWL). Each of these languages is rich enough to express the hierarchical structures inherent in knowledge representation. RDF specializes XML by standardizing meanings for: class, subclass, property, subproperty, domain, range, etc. OWL is a further specialization of RDF; it adds standard meaning for: cardinality, inverse properties, synonyms, and many more concepts in three versions: OWL Lite, Owl DL, and OWL Full. The four languages (RDF, Owl Lite, OWL DL, OWL Full) offer a nested set of language capabilities. We adopted OWL Full due to its anticipated widespread acceptance over the coming years. Our ontologies initially were written in the DARPA Markup Language (DAML), a predecessor to OWL, and converted these ontologies to OWL Full.

OWL has support for numbers only through a W3C specification [8]. This spec defines number types (e.g., real numbers, unsigned integer) and some abilities to create derivations of these types (e.g. the closed interval between 0 and 1). It contains no operations or relations on these numbers. This is a deficiency, because basic scientific concepts are defined in terms of numeric concepts. For example, "brighter", "higher", "later", and "more northerly" are special cases of the "greater than" relation, when applied in specific domains. In particular, spectral regions are defined in terms of wavelength (e.g. visible light is between 0.3 and 0.7 nanometers), atmospheric layers are defined by altitude (e.g. troposphere is between 0 and 15 km), etc. This specification also has no notion of a multidimensional space $\mathbf{R}^n$.

Repositories of OWL ontologies exist to enable the work of others to be extended. However, at present there are no ontologies supporting numeric operations (e.g. "greater than", "max"). Several spatial and temporal ontologies exist, but these ontologies do not exploit the fact that space and time are numerical scales. Therefore, the numerical, space, time, and event ontologies that we develop for SWEET will be submitted to a general OWL ontology library.

General purpose available OWL tools include JAVA and Perl parsers, editors, and visualization tools. These products do not generally support the numerical concepts inherent in the xsd specification.

XML-based languages such as OWL are well suited to data and model exchange, but are less practical for storage and query of large ontologies. Existing database management systems provide the needed functionality in storage and indexing of robust ontologies, including support for data integrity, concurrency control, etc.

Consequently, we adopted the Postgres object-oriented DBMS to store the names and parent-child relations of our ontology elements. We created two-way translators between the internal DBMS representation and the standard XML representation of OWL properties. By placing all term declarations in the DBMS, any search for terms is very rapid.

For representation of spatial concepts, we used bounding polygons to describe regions, where possible. Polygons are a native datatype in Postgres.

VI. DYNAMIC ACCESS TO ONTOLOGY ELEMENTS

Many Earth science facts reside in large external databases. We created OWL wrappers to enable several of these database contents to be accessible as if they were local ontology elements. The databases include three

gazetteers: CIA World Map [9], Getty Thesaurus [10], and the Calle Global Gazetteer [11]. Gazetteers translate vernacular names to and from geographic coordinates. We added polygon boundaries to many gazetteer entries that otherwise contained only rectangular bounding boxes. Also included are the USGS real-time list of earthquakes [12] and the Heavens Above real-time list of satellite locations [13]. An OGC Web Map Server (WMS) [3] import capability was added to acquire images accessible through WMS-compliant servers. A map-based interface demonstrates all of these capabilities by querying the external sources in response to user requests.

VII. Intelligent search engine

A search tool that is aided by an ontology can locate resources without having an exact keyword match. To demonstrate this capability, we created a search tool that consults the SWEET ontology to find synonymous and more specific terms than those requested. The tool then submits the union of these terms to the GCMD search tool and presents the results. The results verified that additional relevant terms were found from the search, relative to the exact keyword search. The search tool is implemented as a web service using the RDF Query Language (RDFQL). Once the synonyms and parent-child relationships have been discovered, the augmented query returns resulting GCMD DIF summaries.

VIII. Future research

Much future research is needed to enable the Semantic Web vision to become a reality. Of particular interest are tools for manipulating and interpreting ontologies. Ontologies "represent knowledge" to the extent that software can semantically interpret the OWL tags. Issues such as these are likely to be addressed by the general ontology community, as they are not specific to the Earth sciences.

The vision of the Semantic Web includes XML tags around technical terms on Web pages to point to the meaning of these terms. It is unclear whether web page developers will take the time to mark up their pages with the appropriate namespaces. An alternative approach, currently under investigation, is to automatically generate the tags during the indexing process. Automatic tag creation involves natural language processing to ascertain the meaning of a term based on its context. In some cases, terms have multiple meanings, and tools such as Latent Semantic Analysis (LSA) [14] can be used to distinguish which meaning was intended, based on the appearance of other associated words in the same document. This investigation will lead to an improved search tool for the Earth Science Information Partner (ESIP) Federation Interactive Network for Discovery (FIND) [15].

Acknowledgment

This work was performed at the Jet Propulsion Laboratory, California Institute of Technology, under a contract with the National Aeronautics and Space Administration. Support was provided by the AIST Program of the NASA Earth Science Technology Office.

References

Fensel, D., J. Hendler, H. Lieberman, W. Wahlster (Eds.), 2003, Spinning the Semantic Web, MIT Press, Cambridge, 479 pp.

Internet references

[1] OWL. http://www.w3.org/TR/owl-ref

[2] DODS. http://www.unidata.ucar.edu/packages/dods

[3] Open GIS Consortium. http://opengis.org

[4] SWEET. http://sweet.jpl.nasa.gov

[5] GCMD Science Keywords and Directory Keywords. http://gcmd.nasa.gov/Resources/valids/index.html

[6] CF Standard name table. http://www.cgd.ucar.edu/cms/eaton/cf-metadata/standard_name.html

[7] Earth Science Markup Language. http://esml.itsc.uah.edu

[8] XML Schema Part 2: Datatypes. http://www.w3.org/TR/xmlschema-2.

[9] CIA World Factbook. http://www.cia.gov/cia/publications/factbook/

[10] Getty Thesaurus of Place Names. http://www.getty.edu/research/conducting_research/vocabularies/tgn/

[11] Calle Global Gazetteer. http://www.calle.com/world

[12] Earthquake List for World. http://earthquake.usgs.gov/recenteqsww/Quakes/quakes_all.html

[13] Heavens Above. http://www.heavens-above.com

[14] Latent Semantic Analysis. http://lsa.colorado.edu.

[15] FIND. http://esipfed.org/find