Semantic Web for Earth and Environmental Terminology (SWEET)

Rob Raskin, Michael Pan

NASA/Jet Propulsion Laboratory
California Institute of Technology
Pasadena, CA 91109
raskin@seanet.jpl.nasa.gov, mjpan@seanet.jpl.nasa.gov

Abstract

In this presentation, we describe our experiences with building and using large ontologies, with application to locating NASA Earth science data. We use OWL to represent the mutual relationships of scientific concepts and their ancillary space, time, and environmental descriptors.

Background

NASA's Earth science mission is to improve our understanding of the integrated Earth system and its components, through the use of satellite data products. NASA makes its data and information products available at no charge to scientists and non-scientists. The motivation of our task is to improve the discovery of these products using tools that incorporate semantic understanding. In support of this effort, we developed a collection of ontologies for describing Earth science data and knowledge. An ontology-aided search tool was developed to demonstrate the use of these ontologies.

Ontologies

The starting point for constructing our Earth science ontologies is the set of keywords in the NASA Global Change Master Directory (GCMD) (Global Change Master Directory, 2003). This collection includes both controlled and uncontrolled keywords.

The controlled keywords include approximately 1000 Earth science terms represented in a subject taxonomy. Several hundred additional controlled keywords are defined for ancillary support, such as: instruments, data centers, missions, etc. The controlled keywords are represented as a taxonomy.

The uncontrolled keywords consist of 20,000 terms submitted by data providers. These terms tend to be more general than or synonymous with the controlled terms. Examples of frequently submitted terms include: climatology, remote sensing, EOSDIS, statistics, marine, geology, vegetation, etc.

SWEET Ontologies

In a taxonomy, properties are not passed on from parent to child, making it less suitable in its present form for knowledge representation purposes. Instead, we used the GCMD keywords as a guide in developing our ontologies, with significant effort devoted to "orthogonalizing" the concept space. For example, "Sea surface temperature" combines a property (temperature), an Earth realm (ocean), and a location (surface). Decomposing science concepts in this way provides a scalable solution to an evolving science knowledge representation environment. The following describes our decomposition of the concept space. We expressed the ontologies using the Ontology Web Language (OWL) promoted by the W3C (World Wide Web Consortium, 2003).

Earth Realm. The "spheres" or environments of the Earth constitute an EarthRealm ontology. Elements of this ontology include "atmosphere", "ocean", and "solid earth", and associated subrealms. The subrealms generally are distinguished from their parent classes, based on the property of altitude. Hence, "troposphere" is the subclass of "atmosphere" where elevation is between 0 and 15 km.

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

Living Element. This ontology includes plant and animal species.

Physical Property. A separate ontology was developed for physical properties that might be associated with any component of EarthRealm, Substance, or Living Element. 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, R^2, etc. Numerical relations include: greaterThan, max, etc.

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

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, inlet, etc. Spatial relations include: above, northOf, etc.

Phenomena. A phenomena ontology is used to define complex processes. A phenomenon crosses bounds of other ontology elements. Examples include: hurricane, earthquake, El Nino, volcano, terrorist event, and each may have associated Time, Space, EarthRealm, NonLivingElement, LivingElement, etc. We include specific instances of phenomena, spanning approximately 50 events over the past two decades.

Human Activities. This ontology is included for representing impacts of environmental phenomena. It includes entries such as: commerce, fisheries, etc.

Ontologies as a Unifying Knowledge Framework

Most of the above ontology categories represent orthogonal concept spaces. Each of these orthogonal dimensions constitutes a hierarchy of complexity (or richness); traversing down the associated tree follows the path of reductionism by adding additional details to more abstract concepts. An additional dimension "phenomena" is synergetic rather than orthogonal to the others. The phenomena entries describe synthesizing concepts that utilize elements from the other ontologies (e.g., a hurricane is associated with particular coastal areas, and is characterized by high winds, rainfall, flood impacts, etc.). Taken together, these complementary dimensions mirror the scientist's dual processes of reductionism and synthesis. This structure provides a semantic framework for classifying resources in terms of their underlying knowledge context.

Numerical Concepts

OWL is limited in its support of numeric concepts, as numbers are supported only through a W3C specification (World Wide Web Consortium, 2001). This spec defines number types (e.g., real numbers, unsigned integer) but makes it very cumbersome to create derivations of these types (e.g., the closed interval between 0 and 1). It contains no operations or relations on these numbers and no notion of a multidimensional space $\mathbf{R}^n$. These are deficiencies for science representation, as many scientific concepts are defined in terms of numerical quantities. For example, spectral regions are defined in terms of wavelength (e.g. visible light is between 0.3 and 0.7 nanometers) and concepts such as "brighter", "higher", "later", or "more northerly" are instances of the "greater than" relation, when applied in specific domains.

Several time and space ontologies already exist. None of these ontologies are expressed as subclasses of numerical scales ($\mathbf{R}^1$ or $\mathbf{R}^3$). Existing space and time ontologies essentially reinvent the numerical scales and their properties. Our approach has been to create numerical ontologies that are used to define space and time concepts.

Spatial Concepts

Using our definitions of two- and three-dimensional space, we adapted a large open-source gazetteer to DAML. We represented regions as polygons, rather than bounding boxes. In DAML, polygons are an extent in a multidimensional space with associated properties (boundary) and associated boundaries (inside, outside, etc.).

Ontology Lessons Learned

From our experience with ontology development, we concluded that the following guiding principles are essential:

Scalability. An ontology should be easily extendable to enable specialized domains to build upon more general ontologies already generated.

Application-independence. The structure and contents of an ontology should be based upon the inherent knowledge of the discipline, rather than on how the domain knowledge is used.

Natural language-independence. The structure should provide a representation of *concepts*, rather than of terms. The concepts remain the same regardless of the inclusion of slang, technical jargon, foreign languages, etc. Synonymous terms (e.g., marine, ocean, sea, oceanography, ocean science) can be mapped separately to an ontology element

Orthogonality. Compound concepts should be decomposed into their component parts, to make it easy to recombine concepts in new ways.

Community involvement. Community input should guide the development of any ontology.

DBMS Storage

XML based languages such as DAML 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.

We have defined a structure of information which is independent of the domain information itself. This ontological structure is compatible with the OWL representation and automatically modifies itself as its contents are updated. To take advantage of existing ontologies, we developed modules to import these ontologies, regardless of their format (XML, RDF, DAML+OIL, OWL, CSV and tab delimited text files).

We adopted the POSTGRES object-oriented DBMS to store the ontology elements. There was no DBMS API available for DAML, so we created two-way translators between the internal DBMS representation and the usual XML representation of the subclass and subproperty relations. By placing all term declarations in the DBMS, searches are very rapid.

PostgreSQL supports geospatial datatypes, such as points, lines, circles, boxes, and polygons, as well as spatial indexing using R-Trees. As such, queries for geospatial boundaries such as those provided in gazetteer are improved by making use of such resources. However, as our generalized ontology structure does not support such resources by default, we build extensions to the default structure while conforming to the essence of the theory. This is done by adding another lookup table for polygon datatype values. This polygon lookup is indexed using an R-Tree index. All queries upon entities within our ontology are processed as before, except when calculating spatial predicates (such as distanceBetween, overlaps, etc.). These spatial predicates then are translated into queries for the PostgreSQL DBMS to resolve.

Ontology-Aided Search

A search tool that is aided by an ontology can potentially locate resources without having an exact keyword match. To verify this claim, we created a search tool that consults the SWEET ontology to find synonymous, less specific, 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.

A search interface to our spatial ontology takes advantage of distributed resources to satisfy requests. When a query requires information about a feature (e.g., the extent of a watershed) that it does not already possess, SWEET is capable of querying existing gazetteers to obtain the required information and update its knowledge base. Human expert knowledge is acquired through a web-based interface which allows users to enter facts and predicates and to update existing knowledge. Our experiences with these components in the spatial domain will help development of other ontologies.

As a further enhancement to ontology-aided search, we explored methods of automatically discovering associations between terms. For example, the terms “carbon dioxide” and “global warming” would likely be high associated, in the sense that when one term appears, the other is relatively likely. We used the GCMD DIF summaries as the text, from which we created an association matrix. We applied latent semantic analysis (LSA) (Berry, 2001), a method that uses empirical orthogonal functions to find additional hidden associations between terms. We included these association scores in the search tool that we developed. The use of DIF summaries probably limits the value of this approach (as the summaries are inconsistent in their content). Nevertheless, this exercise showed that additional relevant associated terms could be automatically extracted. In the next stage of this project, we will examine alternatives to LSA that reduce the large computational requirements of processing large matrices and then apply the methodology to the large corpus of knowledge inherent in the Earth Science Information Partner (ESIP) Federation members’ Web pages.

Future Research

Much future research is needed to enable semantic web tools to become effective. Of particular interest is automation of tasks now being performed manually, including: automatic semantic acquisition, automatic ontology population, and automatic query classification. Accompanying these efforts should be a method of benchmarking, which is necessary to compare our approach with others in the field. There also is a need for better tools for manipulating ontologies. All of these areas are likely to be addressed by the general ontology community, as they are not specific to the sciences.

Acknowledgements

This research was funded by grants from the NASA Earth Science Technology Office and the Earth Science Information Partner (ESIP) Federation. This work was carried out at the Jet Propulsion Laboratory of the California Institute of Technology.

References

Berry, M., ed. 2001. *Computational Information Retrieval*. Philadelphia, SIAM.

Global Change Master Directory 2003. *GCMD’s Earth Science Keywords*, http://gcmd.nasa.gov/Resources/valids.

World Wide Web Consortium 2001. *XML Schema Part 2: Datatypes*, http://www.w3.org/TR/xmlschema-2.

World Wide Web Consortium 2003. *OWL Reference*
http://www.w3.org/TR/2003/CR-owl-ref-20030818.