Computers & Geosciences 31 (2005) 1119–1125

COMPUTERS & GEOSCIENCES

www.elsevier.com/locate/cageo

Knowledge representation in the semantic web for Earth and environmental terminology (SWEET)

Robert G. Raskin*, Michael J. Pan

Jet Propulsion Laboratory, Mail Code 300-320, California Institute of Technology, Pasadena, CA 91109, USA

Received 14 May 2004; received in revised form 6 October 2004; accepted 31 December 2004

Abstract

The semantic web for Earth and environmental terminology (SWEET) is an investigation in improving discovery and use of Earth science data, through software understanding of the semantics of web resources. Semantic understanding is enabled through the use of ontologies, or formal representations of technical concepts and their interrelations in a form that supports domain knowledge. The ultimate vision of the semantic web consists of web pages with XML namespace tags around terms, enabling search tools to ascertain their meanings by following the link to the defining ontologies. Such a scenario both reduces the number of false hits (where a search returns alternative, unintended meanings of a term) and increases the number of successful hits (where searcher and information provider have a syntax mismatch of the same concept). For SWEET, we developed a collection of ontologies using the web ontology language (OWL) that include both orthogonal concepts (space, time, Earth realms, physical quantities, etc.) and integrative science knowledge concepts (phenomena, events, etc.). This paper describes the development of a knowledge space for Earth system science and related concepts (such as data properties). Some of the ontology contents are "virtual" by means of an OWL wrapper associated with terms in large external databases (including gazetteers and Earthquake databases). We developed a search tool that finds alternative search terms (based on the semantics) and redirects the expanded set of terms to a search engine.

Keywords: Ontology; Classification; Semantics; Search; Discovery

1. Introduction

Web searches for Earth science data and information are commonly hindered by syntax mismatches between information user and information provider. If the user does not enter the "correct" search terms, either not enough or too many hits are returned. The underlying cause is that the search tool is unable to extract meaning from terms on web pages. An emerging solution to this problem is through the "semantic web" (Berners-Lee et al., 2001), an ambitious extension to the existing WWW environment proposed by the world wide consortium (W3C). In a semantic web, words appearing in web resources are linked to corresponding entries in ontologies, where terms are defined and their mutual relationships are clarified.

The ultimate objective of our task is to improve semantic understanding of web resources by software tools, with particular application to discovery and use of Earth science data. Semantic understanding of text by automated tools is enabled through the combination of (i) ontologies and (ii) software tools that can interpret the ontologies and apply any needed reasoning. An

*Corresponding author. Fax: +1 818 3932718.
E-mail address: raskin@seastar.jpl.nasa.gov (R.G. Raskin).

0098-3004/$ - see front matter © 2005 Elsevier Ltd. All rights reserved.
doi:10.1016/j.cageo.2004.12.004

ontology (in the computer science sense) 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 the parent concept(s). An ontology is crucial for describing the semantic content of data, to complement the syntactic content that appear in Earth science markup language (ESML) (Ramachandran et al., 2004) descriptor files or other metadata descriptions.

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 can also locate resources without having an exact keyword match; e.g., "El Nino" has an equivalent definition in terms of physical properties of the tropical Pacific Ocean.

To support potential semantic web activities, we developed a collection of ontologies for the Earth and environmental sciences and supporting areas, as part of the semantic web for Earth and environmental terminology (SWEET) project. The ontologies are written in the web ontology language (OWL) (Dean and Schreiber, 2004), an XML-based standard adopted by the W3C, and serve as a common sense knowledge base of Earth system science concepts. We used these ontologies in a prototype search tool that improves performance by creating additional relevant search terms based on the underlying semantics. The remainder of this paper focuses on the ontology development rather than specific semantic web applications or automated reasoning algorithms.

2. Ontology representation

HTML documents include tags such as ⟨b⟩ and ⟨/b⟩ to request web browsers (or other HTML-aware software) to start and end boldface presentation. XML documents contain tags with arbitrary names, so there is no such shared understanding of tag meanings. In this sense, XML provides a specification for syntax but not semantics. Specialized XML languages have emerged to predefine several standard terms to provide a starting point for shared understanding of concepts. The resource description framework (RDF) is the simplest such language; it defines standard meanings for: class, subclassOf, property, subpropertyOf, domain, range, sequences, collections, and a few other terms. RDF-aware software tools understand that a ⟨*subclassOf*⟩ ⟨/*subclassOf*⟩ pair indicates that one concept is a specialization of another. This structure provides the infrastructure to create simple ontologies which include hierarchical semantic relationships of arbitrary depth and multiple inheritance.

OWL is a further specialization of RDF; it includes most RDF definitions and adds standard meaning for functional properties, inverse relations, transitivity, synonyms, cardinality restrictions, and additional concepts. OWL itself has three versions of increasing complexity: OWL Lite, Owl DL, and OWL Full. OWL DL adds to OWL Lite the concepts of Boolean and set complement relations. OWL DL is named for its alignment with *descriptive logic*. Its rules support computational completeness (all conclusions are guaranteed to be computable) and decidability (all computations will finish in finite time). OWL Full uses the same language components as OWL DL, but relaxes restrictions on their use (e.g., OWL Full enables a class to be an instance of another class). Most software tools do not support OWL Full due to the time required to validate queries. SWEET is currently consistent with the OWL DL specification.

Both RDF and OWL have been accepted as ontology languages by the W3C. This standardization ensures that software tools will support these languages in the near future, and enables RDF or OWL users to "import" (or extend) ontologies created by others. Rather than creating one single ontology for the English language (as had been initially sought) (Lenat and Guha, 1990), developers need only create extensions to existing upper-level ontologies that fit their domain needs. OWL has its limitations, e.g., it has no concept of variables; however, no competing ontology languages remain in widespread use.

3. 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 that specialize their parent concept(s). Thus, *surface water* has the child concept of *river*, which itself has *Mississippi River* as an instance. So, *river* inherits all properties of *surface water* (with added attributes such as being *inland* and having associated *outlet*) and Mississippi inherits all properties of *river* (with additional location and size properties). In this paper, we describe our experiences with the development of Earth and environmental science ontologies. Our starting point of reference was the collection of keywords in the NASA global change master directory (GCMD) (Olsen, 2001). This collection includes both controlled and uncontrolled keywords.

Controlled keywords: GCMD includes approximately 1000 controlled Earth science keywords, represented as a *taxonomy*. A taxonomy is a *subject* classification, as used by libraries or clearinghouses to classify resources. In a taxonomy, properties are generally not passed down from parent to child, making this structure less suitable for knowledge representation purposes. Several hundred additional *non-science* terms reside in the GCMD controlled keyword set to describe dataset attributes, data services, instruments, data centers, and missions, etc. Relevant NASA-funded data collection activities are required to submit metadata records to the GCMD, using keywords designated by the data provider.

Uncontrolled keywords: Data providers may submit additional *uncontrolled keywords* within their metadata records, and over 20,000 such terms have been submitted to date. Many of these terms are more abstract than the controlled keywords, frequently submitted uncontrolled terms include: climatology, remote sensing, EOSDIS, statistics, marine, geology, and vegetation. Other submitted uncontrolled terms are narrow specializations (or synonyms) of controlled keywords.

The GCMD controlled keywords provided an initial guide in developing our ontologies, although several structural changes were made to the keyword structure. Rather than define a compound concept such as *air temperature*, we separated the physical property (*temperature*) from the element that the property applies to (*air*). This provides a more scalable solution to a growing knowledge base. In this case, knowledge of the independent concepts of the substance *air* and property *temperature* provide a complete understanding of *air temperature* without a need to create an explicit definition of the compound concept. Such a decomposition does not preclude term recomposition by Earth system science communities, but the compound term would be declared as being synonomous with its component parts (*substance* = *air and property* = *temperature*). In some cases, the compound concepts contain more meaning than their component parts (e.g., *static pressure*) and the compound term is explicitly included.

Additional sources of terms include the uncontrolled GCMD keyword collection, as well as other controlled science lists, such as the CF Standard Names list (Eaton et al., 2003). The CF vocabulary consists of over 500 terms, many of which relate to complex relationships useful to the modeling communities, such as a parameter under given physical conditions. Such terms decompose readily into SWEET components. Given all of these sources, we manually populated the ontologies using the OilEd, Protégé, and Construct ontology editors.

The resulting ontology set is intended to be a concept space rather than a controlled vocabulary. As an upper-level ontology for Earth system science, we anticipate that subdisciplines will combine existing terms or add new ones to satisfy specialized domain needs. From our experience with ontology development, we concluded that the following guiding principles are essential:

1. *Scalability*: An ontology should be easily extendable to enable specialized domains to build upon more general ontologies already generated.
2. *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.
3. *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.
4. *Orthogonality*: Compound concepts should be decomposed into their component parts, to make it easy to recombine concepts in new ways.
5. *Community involvement*: Community input should guide the development of any ontology.

Our resulting ontologies adhere closely to the above principles, although community involvement is the most challenging goal, and remains an ongoing process.

In addition to these generic principles, we sought any design that could capture the spirit and progress of Earth system science. Scientific advancements often occur via reductionism, where specialists decompose entities (e.g., atoms, genes) into component parts (Pap, 1962; Kuhn, 1962). RDF includes the "subclass" relation, but "part of" is not predefined in RDF or OWL. Some argue for a central role for "part of" in ontologies (Smith, 1998), but at present, the concept must be defined by OWL users as a transitive function (with a corresponding inverse function "has a part").

The opposite of reductionism can be called *holism*, a synergetic concept popularized by general system theorists (Bertalanffy, 1969). Holism is a common theme in Earth system science, as it describes macroscopic phenomena that may require knowledge from multiple disciplines (Schneider and Boston, 1991). Integrative concepts can be described in OWL using the "has a part" function (inverse of "part of") and "has associated" relations.

4. SWEET ontologies

The SWEET ontologies include several thousand terms, spanning a broad extent of Earth system science and related concepts (such as data characteristics). To support such a large collection and adhere to the guiding principles, the concepts are divided, where possible, into orthogonal dimensions or *facets* in support of reductionism. The primary ontologies are shown in Fig. 1 and described in Section 4.1. Each box represents a separate ontology, and a connecting line indicates where major *properties* are used to define concepts across ontology spaces. The unifying ontologies (such as phenomena) are generally holistic, in derivation. For example, a hurricane is associated with particular coastal areas, and is characterized by high winds, rainfall, flood impacts, etc.

Fig. 2 shows an excerpt from the Phenomena ontology. *StormSurge* is defined as a special case of (*subclassOf*) *Flood*, where the *EarthRealm* has the value of *CoastalRegion. Flood* is itself defined as a special case of *SevereWeatherPhenomena* with hasAssociatedSubstance of *LiquidWater* and hasAssociatedEarthRealm of *LandSurface.* The *hasAssociated* relations represent many of the connecting lines between ontologies in Fig. 1.

4.1. Ontology descriptions

EarthRealm: 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 are often 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. This ontology can be considered a ‘‘state’’ of the planet that is extendable to past or future time periods, (as well as to other planets).

NonLivingSubstances: The non-living building blocks of nature include: particles, electromagnetic radiation,

```
<owl:Class rdf:ID="StormSurge">
  <rdfs:subClassOf rdf:resource="#Flood"/>
  <rdfs:subClassOf>
    <owl:Restriction>
      <owl:onProperty rdf:resource="#hasAssociatedEarthRealm"/>
      <owl:allValuesFrom rdf:resource="#earthrealm.owl#CoastalRegion"/>
    </owl:Restriction>
  </rdfs:subClassOf>
</owl:Class>


<owl:Class rdf:ID="Flood">
  <rdfs:subClassOf rdf:resource="#SevereWeatherPhenomena"/>
  <rdfs:subClassOf>
    <owl:Restriction>
      <owl:onProperty rdf:resource="#hasAssociatedEarthRealm"/>
      <owl:allValuesFrom rdf:resource="earthrealm.owl#LandSurface"/>
    </owl:Restriction>
  </rdfs:subClassOf>
  <rdfs:subClassOf>
    <owl:Restriction>
      <owl:onProperty rdf:resource="#hasAssociatedSubstance"/>
      <owl:allValuesFrom rdf:resource="substance.owl#LiquidWater"/>
    </owl:Restriction>
  </rdfs:subClassOf>
</owl:Class>
```

Fig. 2. Sample SWEET excerpt from phenomena ontology.

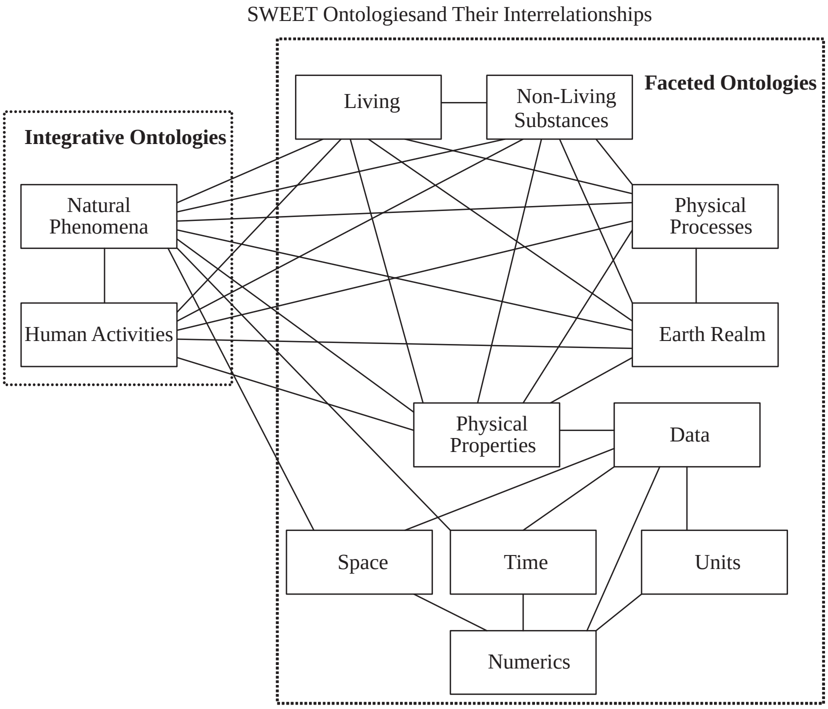

Fig. 1. SWEET ontologies and their interrelationship.

and chemical compounds. These substances constitute an ontology of physics and chemistry.

LivingSubstances: The living substances include plant and animal species. This ontology was imported from the "biosphere" taxonomy of GCMD.

PhysicalProcesses: Physical processes include processes that affect living and non-living substances, such as diffusion, evaporation, etc.

PhysicalProperties: A separate ontology was developed for physical properties, including those observable or associated with other components. *PhysicalProperties* include "temperature", "pressure", "height", "composition", etc., and could apply to *NonLivingSubstances*, *LivingSubstances*, *PhysicalProcesses*, etc. These properties typically are measured physical quantities (or qualities) with units.

Units: Units are defined using Unidata's UDUnits (Unidata, 1997). 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.

Time: 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.

Space: Space is essentially a multidimensional 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.

Numerics: Numerical extents include: interval, point, 0, $\mathbf{R}^2$, etc. Numerical relations include: greaterThan, max. etc. We defined multidimensional concepts, as these are not native to the OWL and XML environments.

PhysicalPhenomena: 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 phenomena, spanning approximately 50 events over the past two decades.

HumanActivities: This ontology is included for representing activities that humans engage in, such as commerce, fisheries, etc. This ontology is included because scientific processes and phenomena have human impacts, and there is a need for representing such activities.

Data: The data ontology provides support for dataset concepts, including representation, storage, modeling, format, resources, services, and distribution.

5. Issues and applications

5.1. Numerical concepts

A deficiency of RDF and OWL is that the languages contain no direct support for numerical concepts, and must rely instead on a limited XSD (XML Schema Definition) specification of datatypes (Byron and Malhotra, 2001). This spec defines number types (e.g., floating point, unsigned integer) and methods to create derivations of these types (e.g. the closed interval between 0 and 1), but contains no operations or relations on these numbers. This is a deficiency, because many scientific concepts are defined through numeric concepts. For example, "brighter", "higher", "later", and "more northerly" are special cases of the "greater than" relation, applied in specific domains. In particular, spectral regions are defined in terms of wavelength (e.g., visible light is between 0.4 and 0.7 μm), 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 $\mathbf{R}^n$. The *Numerics* ontology adds extensions needed to define scientific concepts and is used to define concepts in the spatial and temporal ontologies. Although other spatial and temporal ontologies exist, none exploit the fact that space and time are numerical scales. Without such a connecting thread, many numerical concepts must be reinvented to create definitions of space and time.

5.2. Storage of ontology elements

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. We anticipate that native OWL DBMS software will emerge to efficiently handle OWL ontologies, and our implementation in this area should be considered only a temporary solution.

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 now, both the XML and DBMS representations are actively used. For search, the DBMS approach is optimal, while for reasoning applications, the XML representation is more efficient. For representation of spatial concepts, we used bounding polygons to describe regions, where possible; polygons are a native datatype in Postgres.

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 contents of two gazetteers: CIA World Map (CIA, 2004) and Getty Thesaurus (Getty, 2004) have been incorporated, as well as the USGS list of Earthquakes (USGS, 2004).

5.3. Ontology-aided search

The single application to date of SWEET is to aid search of Earth science data resources. A search tool that can interpret RDF or OWL ontologies can potentially locate resources without the need for an exact keyword match. To demonstrate this capability, we created a search tool using the Perl language that consults the SWEET ontology and finds synonymous, more specific terms, and more general terms than those requested. The union of these terms is submitted to the GCMD search tool and the results are presented. This infrastructure is currently being implemented in the search mechanism of the Earth science information partner (ESIP) Federation (Raskin et al., 2002). The ESIP Federation is a consortium of 75 data and service providers spanning many subdisciplines within Earth system science. In the implementation, the search results are returned as clusters, with synonymous, parent, and child matches presented separately.

As an example, a search for "marine temperature" found resources using the following alternate definitions for marine: *Marine ecosystem*, *submarine canyon*, *trench*, *marine wetland*, *wetland*, *sediment*, *ocean layer*, *deep ocean*, *ocean crust layer*, etc.; and for temperature: *equivalent temperature*, *virtual temperature*, *brightness temperature*, etc.

6. Conclusions and future research

The primary purpose of the SWEET project was to create a knowledge base, rather than to create applications of the knowledge. We envision many applications to emerge, as general-purpose semantic web tools became more widespread. Other ontologies for the Earth sciences exist, but are limited to specific subsets of the Earth system. For example, the GEON ontology (Seber et al., 2003) is limited to the solid Earth. Collaboration is underway to maintain consistency between the SWEET and GEON ontologies.

Additional research is needed to enable the semantic web vision to become an operational reality. Of particular interest are 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 is also a need for better tools for manipulating ontologies. Most of these areas are likely to be addressed by the general ontology community, as they are not specific to the Earth sciences.

The semantic web vision consists of XML tags placed around technical terms on web pages, which point to the term meanings. It is unclear whether web page developers will take the time to mark up their pages with links to the defining namespaces. An alternative approach that we are currently investigating 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) (Landauer et al., 1998) could be used to distinguish which meaning was intended, based on the appearance of other associated words in the same document.

Acknowledgements

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

Berners-Lee, T., Hendler, J., Lassila, O., 2001. The semantic web. Scientific American, 28–37.

Bertalanffy, L., 1969. General System Theory: Foundations, Development, Applications. George Braziller, Inc., New York, NY 290pp.

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

CIA, 2004. CIA World Factbook. http://www.cia.gov/cia/publications/factbook/.

Dean, M., Schreiber, G., 2004. OWL Web Ontology Language Reference, W3C. http://www.w3.org/TR/owl-ref.

Eaton, B., Gregory, J., Drach, B., Taylor, K., Hankin, S., 2003. NetCDF Climate and Forecast (CF) Metadata Conventions Version 1.0, University Consortium for Atmospheric Research (UCAR), 23 October 2003. http://www.cgd.ucar.edu/cms/eaton/cf-metadata/CF-1.0.html.

Getty Trust, 2004. Getty Thesaurus of Place Names. http://www.getty.edu/research/conducting_research/vocabularies/tgn/.

Kuhn, T.S., 1962. The Structure of Scientific Revolutions. University of Chicago Press, Chicago, IL 226pp.

Landauer, T.K., Foltz, P.W., Laham, D., 1998. Introduction to latent semantic analysis. Discourse Processes 25, 259–284.

Lenat, D.B., Guha, R.V., 1990. Building Large Knowledge-based Systems: Representation and Inference in the Cyc Project. Addison-Wesley, Reading, MA 372pp.

Olsen, L., 2001. Supporting interagency collaboration through the Global Change Data and Information System (GCDIS). Proceedings of the American Meteorological Society, Long Beach, CA, January 2001, pp. 348–351.

Pap, A., 1962. An Introduction to the Philosophy of Science. The Free Press, New York, NY 292pp.

Ramachandran, R., Graves, S., Conover, H., Moe, K., 2004. Earth Science Markup Language (ESML): a solution for scientific data-application interoperability problems. Computers & Geosciences 30 (1), 117–124.

Raskin, R., Burrows, H., Conover, H., Gallagher, J., Major, G., Rhyne, T., 2002. Discovering and accessing data from the Earth Science Information Partner (ESIP) Federation. EOS 83 (47), 543.

Schneider, S.H., Boston, P.J. (Eds.), 1991. Scientists on Gaia. MIT Press, Cambridge, MA 433pp.

Seber, D., Keller, R., Sinha, K., Baru, C., 2003. GEON: cyberinfrastructure for the geosciences. EOS Transactions American Geophysical Union, 84(46), Fall Meeting Supplement, Abstract U21A-04.

Smith, B., 1998. The basic tools of formal ontology. In: Gualino, N. (Ed.), Formal ontology in information systems. IOS Press, Amsterdam, The Netherlands 347pp.

Unidata, 1997. UD units: a library for manipulating units of physical quantities. Unidata, Boulder, CO. http://my.unidata.ucar.edu/content/software/udnits.

USGS, 2004. Earthquake list for world. http://Earthquake.usgs.gov/recenteqsww/Quakes/quakes_all.html.