Ontology based Semantic Metadata for Geoscience Data

Viral Parekh
Department of Computer Science and Electrical Engineering,
University of Maryland, Baltimore County
Baltimore, MD 21250
Email: viral1@umbc.edu

Jin-Ping Gwo
Department of Civil and Environmental Engineering
University of Maryland, Baltimore County
Baltimore, MD 21250
Email: jgwo@umbc.edu

Tim Finin
Department of Computer Science and Electrical Engineering,
University of Maryland, Baltimore County
Baltimore, MD 21250
Email: finin@umbc.edu

Abstract

In Geoscience domain, large amounts of data are accessible, however they vary in formats and are stored at various organizations leading to problems of data discovery, data interoperability and usability. In this paper, we propose a new semantic metadata paradigm based on ontologies and the use of Semantic Web languages. Our suggested data model ontology is used to guide the generation of metadata for individual datasets. This data model ontology defines elements to incorporate information about data identification, spatial extent, temporal extent, data presentation form, data content and data distribution regarding the dataset. Combining domain specific ontologies with this data model ontology offers a new approach to the generation of semantic metadata for datasets. The system allows the data provider to select concepts from domain ontologies that best describe the content within the dataset. This selection along with the links to domain ontologies is stored within the metadata file, thereby generating semantic metadata for the dataset. This metadata is capable of facilitating the end users of data with content based discovery of datasets irrespective of their locations and formats.

Keywords

ontology, data discovery, metadata schema, semantic web

1. Introduction

Huge volumes of Geoscience data are available and accessible to researchers all over the world. There are several data providers such as US Government agencies like Environmental Protection Agency (EPA), United States Geological Survey (USGS), National Oceanic & Atmospheric Administration (NOAA), National Aeronautics and Space Administration (NASA), etc and other non-profit organizations like National Center for Atmospheric Research (NCAR). They produce different kinds of data which is archived at various locations and distributed in many different formats. This variety of formats leads to data interoperability and data usability problems faced by the researchers and other users. Also, the datasets are distributed and stored by various organizations making the task of locating and retrieving the relevant datasets very complex. There is a vital need of an efficient mechanism for discovery of required datasets. The end users of these geoscience datasets could be researchers searching for relevant data to perform certain experiments or modeling tasks, people from industries looking for right data in order to facilitate decision making or even students in search of data for their class projects.

In this paper, we propose a semantic metadata management system based on ontologies and use of Semantic Web languages. This proposed system will address the data discovery problem and provide a basis for data interoperability and usability. The objective of this system is to provide a metadata paradigm that is semantically rich and capable of facilitating content based discovery of datasets to the end users, irrespective of the formats and locations of the datasets. Our ultimate vision is to build intelligent and powerful environmental information systems by developing information

infrastructures that may enable the deployment of efficient data sharing and integration mechanisms. We see our current work in building ontology based semantic metadata management system as a first step towards our final objective of semantic interoperability.

FGDC (Federal Geographic Data Committee) Content Standard for Digital Geospatial Metadata [9] was developed in 1994 to describe all possible geospatial data. However, the standard is very complex with 334 different elements, 119 of which exist only to contain other elements making this standard difficult to use. Moreover, the standard provides text based syntactic metadata with virtually no semantics and machine understandability when compared to the proposed ontology based semantic metadata.

Ontologies are designed to provide an abstract conceptualization of information and a vocabulary of terms to be used in this representation. They provide semantics to the domain and define the set of domain concepts and relationships among these concepts. This paper talks about our approach in using a set of ontologies to provide semantic metadata for datasets compared to the traditional approach of using text based syntactic metadata. The motivating factors for using ontology based approach for generating semantic metadata schemas are:

- Ontologies can be constructed to provide a shared, common vocabulary involved in describing the dataset, thereby defining a standard of metadata which can be used by all.
- Ontologies can provide a conceptual schema for any dataset regardless of its format, structure or size.
- Ontologies can be designed to semantically understand the content and structure of data present in the dataset.
- Ontologies can be used to help the data providers to enter the metadata in a semantically valid form.
- Interoperability among heterogeneous datasets can be achieved by using shared ontologies.
- Ontologies are viewed as the most advanced knowledge representation model.
- Ontology can be used as a basis for content based discovery and retrieval of datasets.

We have encoded the ontologies in Web Ontology Language (OWL) [2], a W3C recommendation that is designed to realize the Semantic Web. The Semantic Web is a future vision in which information is given well defined meaning using ontologies, thereby enabling the machines to understand and process the available information [1]. The Semantic Web and OWL are designed for extending syntactic interoperability to semantic interoperability. OWL provides extensive vocabulary along with formal semantics and facilitates machine interpretability. The expressive power of OWL adds more semantics to our ontologies. The semantic metadata generated by using these OWL ontologies are encoded as OWL files and hence machine understandable and also available to the future Semantic Web.

In section 2 we discuss our methodology of registration of datasets by the data providers and the generation of semantic metadata. Section 3 describes in detail the data model ontology and its different components. This data model ontology defines the vocabulary required for generating the semantic metadata. We briefly discuss and compare some of the related work in Section 4. Section 5 concludes our paper.

2. Dataset Registration

Figure 1 depicts the complete dataset registration process. The role of data providers is to register their datasets using a semantically valid form which in turn uses a set of ontologies. This registration process generates semantic metadata for the dataset which is stored in the knowledge base.

As can be seen, the ontology repository consists of several ontologies in OWL – the data model ontology and other domain ontologies such as geoscience, spatial and temporal ontologies. The data model ontology contains defined classes and properties to facilitate the creation of

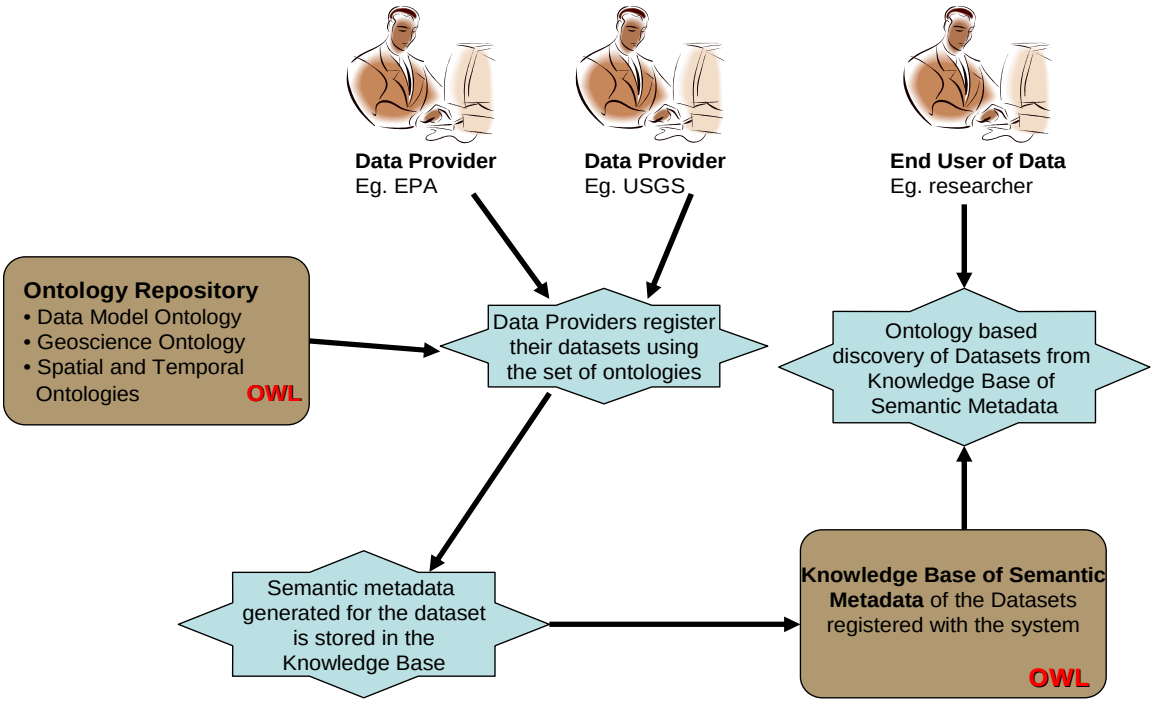

Figure 1 Data Registration and Semantic Metadata Generation

metadata for the dataset. It also includes provisions to incorporate semantic understanding of the dataset content within the metadata. This semantic understanding is achieved by the use of geoscience, spatial and temporal ontologies which define all the required domain concepts and the relationships among them. Semantic elements from these domain ontologies are embedded within the metadata files along with links to the ontologies where they are defined. By the inclusion of these semantic elements corresponding to the data fields within the dataset, semantic metadata for the dataset is generated. The knowledge base stores the semantic metadata of individual datasets registered with the system. This semantic metadata is an OWL instance file of the data model OWL ontology. The knowledge base is hence a collection of OWL files, one for each dataset. The end user who is in need of data such as a researcher could then query this knowledge base of semantic metadata in order to fetch the relevant datasets.

3. Data Model Ontology

The data model ontology facilitates the registration of datasets by the data providers. It provides a standard vocabulary of terms to be used. It also provides end users of data with a mechanism to query for relevant datasets. In this paper, we will focus on the use of data model ontology to facilitate dataset registration.

The objective of this ontology is to provide metadata for the dataset as well as to provide a semantic understanding of the data content within the dataset. The ontology defines a set of elements which will be used for the purpose of documentation of the dataset. It answers who, what, why, where, when and how of every facet of the dataset. The ultimate goal is to provide a basis for an efficient mechanism of content based retrieval of datasets.

Semantic understanding is achieved by mapping the dataset to concepts defined in the geoscience domain ontology. This mapping provides ontology based conceptual schema for the dataset. Data model ontology is designed to provide this connection of the dataset to the Geoscience ontology and in this process a semantic representation of the dataset is generated.

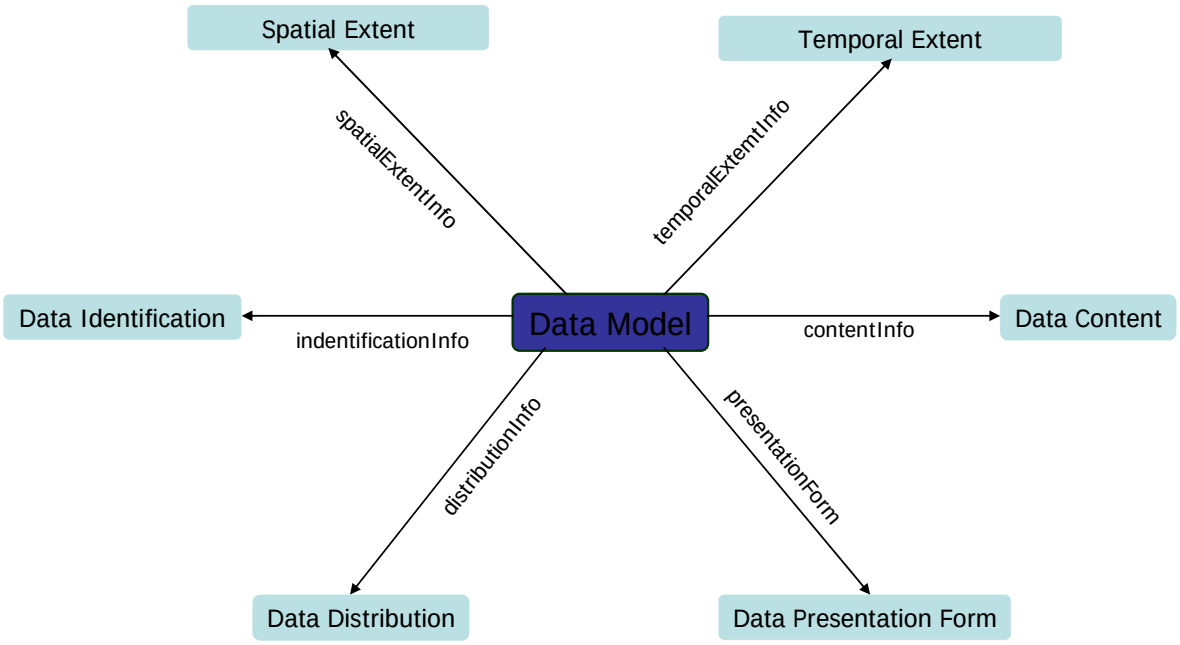

Figure 2 Data Model Ontology

Figure 2 gives an overview of the data model ontology. As can be seen, several classes constitute the data model ontology. The data model ontology is available at http://www.cs.umbc.edu/~viral1/ontologies/datamodel.owl. A brief description of the different classes involved in this ontology follows:

Data Model: Data model class is the principal component of the ontology and it links to other classes in the ontology through its attributes as can be seen in the figure. For each data set that is registered with our system, a corresponding instance of data model class is created and stored in the knowledge base.

Data Identification: This class allows the provider to specify basic identification information about the dataset. The important attributes of this class are:

- *title, description, associatedPublication*
- *creator, participant, pointOfContact*
- *creationDate, lastModificationDate*
- *statusCode, maintenanceFrequency*

This class uses Person and Publication ontologies developed by the ebiquity research group [8] of UMBC. Many attributes of this class are sub-properties of Dublin Core [7] metadata element set (OWL version) which provide a standard for information resource description. Certain attributes such as *statusCode* and *maintenanceFrequency* have an enumeration of allowable values.

Spatial Extent: This class gives information about the geographic area covered by the dataset. It permits the data provider to specify the bounding coordinates of coverage of the dataset in terms of latitude and longitude values in the order western-most, eastern-most, northern-most, and southern-most.

Temporal Extent: This class provides a means for stating the temporal information corresponding to the dataset. It is possible to specify a single date, multiple dates or a range of dates.

Data Presentation Form: Information about form of dataset, i.e. whether it is digital or exists in hardcopy is provided using this class in the ontology. It is also capable to convey whether the dataset is a map, table, document, image, video, profile or model.

Data Content: This is a pivotal class in the data model ontology and is responsible for mapping the dataset to the domain concepts

defined in the Geoscience ontology. This linkage generates a semantic conceptual schema for the dataset. The data provider selects the concepts from the Geoscience ontology that best describe the dataset. This selection is stored in the data content class allowing the data model ontology to provide not only metadata about the dataset but also semantic description of the data content within the dataset.

Data Distribution: Information about the distributor of the dataset and the digital transfer options for obtaining this dataset from the concerned organization can be provided using this class. It also has provisions to specify any legal disclaimer and any use or access constraints associated with the dataset. Also, the software used to access the data could be specified using this class.

4. Discussion

When compared to the traditional metadata standard of FGDC, we believe our metadata standard is simple yet resourceful, semantically rich and machine understandable as it is based on domain rich ontologies which are encoded in OWL. It facilitates ontology based querying for datasets compared to keyword searches for FGDC metadata files. We believe that requiring data providers to register with our system and publish metadata files will not be a burden for them when compared to the relatively large and complex FGDC metadata files they require to create for their datasets. Also, the gains are abundant. Islam A. et al [5] are developing a metadata ontology based on FGDC metadata standard making it very complex and difficult to use as compared to our ontology.

There are other on-going projects in using Semantic Web technologies to improve data discovery, usability and interoperability. As a part of Semantic Web for Earth and Environmental Terminology (SWEET) [4] project at NASA, they have developed several domain ontologies to describe earth science data and knowledge. Their motivation is to improve the discovery of NASA information and data products. In earthquake science community, [6] proposes to develop a data semantics based system to improve interoperability among heterogeneous earthquake data. Also, the Earth System Grid (ESG) project [3] aims to provide discovery of large datasets based on grid technologies and the use of metadata schemas and prototype ontology. However, none of them strive to develop to a semantic metadata standard that can be used by everyone. Moreover, our use of Web Ontology Language (OWL) provides more semantic power to the metadata and also makes the semantic metadata files available to the next generation Semantic Web.

5. Conclusion

In this paper, we discussed our data model ontology and the mechanism of generating ontology based semantic metadata for datasets. Each dataset that is registered with our OWL ontologies has content based semantic description associated with it apart from the metadata information about identification, spatial, extent, distribution and presentation form. This semantic description is independent of the dataset format and is generated using the geoscience domain specific ontologies. This approach allows the end users of data to search for relevant datasets based on their semantic content and metadata rather than just simple keywords. We argue that similar approach of metadata standard would be beneficial to other domains such as geophysics, chemistry, etc if adopted by them as these domains face similar problems of data heterogeneity, data usability and relevant data discovery as faced by the geoscience domain.

6. References

[1]Berners-Lee T., Hendler J., Lassila O., The Semantic Web, Scientific American, May 2001

[2]W3C: OWL Web Ontology Language Semantics and Abstract Syntax.
URL:http://www.w3.org/TR/2004/REC-owl-semantics-20040210/, 2004.

[3]Pouchard L. et al, The Earth System Grid Discovery and Semantic Web Technologies,

Workshop on Semantic Web Technologies for Searching and Retrieving Scientific Data, Florida, USA, 2003

[4] Raskin R., Pan M., Semantic Web for Earth and Environmental Terminology (SWEET), Workshop on Semantic Web Technologies for Searching and Retrieving Scientific Data, Florida, USA, 2003

[5] Islam A. et al, Ontology for Geographic Information - Metadata (ISO 19115),
URL: http://loki.cae.drexel.edu/~wbs/ontology/

[6] Chen A. et al, Interoperability and Semantics for Heterogeneous Earthquake Science Data, , Workshop on Semantic Web Technologies for Searching and Retrieving Scientific Data, Florida, USA, 2003

[7] The Dublin Core Metadata Initiative
URL: http://www.dublincore.org/

[8] The Ebiquity Research Group
URL: http://ebiquity.umbc.edu/

[9] FGDC Metadata
URL: http://www.fgdc.gov/metadata/metadata.html