Semantic Registration of Geoscientific Data through ESIP Semantic Web Testbed

Wenwen Li[1], Chaowei Yang[1] and Rob Raskin[2]

[1]. Center for Intelligent Spatial Computing, George Mason University, Fairfax, VA

[2]. Jet Propulsion Laboratory, NASA, Pasadena, CA

1. Introduction

Massive volumes of earth science data have been generated and archived for decades. However, most of the produced data remain "isolated information islands" due to the absence of efficient data sharing infrastructure and lack of common language that researchers can use to interpret each other's data (Rezgui et al. 2007). As a result, the ability to discover, access, and integrate these large data repositories is still limited. To overcome this drawback, standards for metadata schema registries have been conducted over the past decade. The historical line can be traced back to shared data dictionaries and registration process in the light of ISO/IEC-1 11179 specification (1999). It specifies best practice for data element definition in enabling the registration process. OGC catalogue service is the most popular standard which supports the interoperable registration of geospatial data and services, aiming to supports the ability for data providers to register metadata that conform to a certain information model, e.g. ISO metadata profile (Voges and Senkler, 2007) and ebRIM profile (Martell, 2009). The important benefit of this approach is to increase the reuse of existing metadata and avoid duplication of effort by merging schemas to one view.

However, there are still several problems exacerbated in cases of syntactic and semantic data difference (Zaki et al. 2007). Previous work focused on the syntax-level interoperability, which incumbers effective discovery and integration in semantic level. The project of ESIP/FGDC Semantic Web Testbed aims to enhance the semantic integration of geospatial resources by: (1) addressing the challenge in designing a substantial Knowledge Base (KB) for resource integration and make it available to a broad audience in Earth Science community; (2) developing an online interface and cloud infrastructure to allow semantic registration of datasets and other web resources; (3) provide SPARQL capability to validate backbone ontology and the registration process; (4) support service chaining for spatial decision making. ESIP datatype and service ontology are used as the supporting semantic schema.

2. A Use Case

The use case described below was developed by domain scientists with assistance from the project team to reflect the particular science areas of interest to the geoscientific dataset, service and workflow.

"Generate the annual rainfall data of VA by subsetting the global precipitation dataset distributed from NetCDF server and visualize it in JPEG format".

To address this use case, we need a registry that contains (1) Datasets: Global precipitation data in NetCDF format; (2) one subsetting service and one format conversion service (from NetCDF to JPEG); as well as (3) semantic reasoning capability that can handle the SPARQL queries.

3. System Architecture

We aim to provide a collaborative development testbed to support the use case. Figure 1 presents a scenario for collaboratively building the repository for resource integration. Any user who has internet access can login in to the system to semantically register the scientific content. The main components include services for ontology registration, query, reasoning and access.

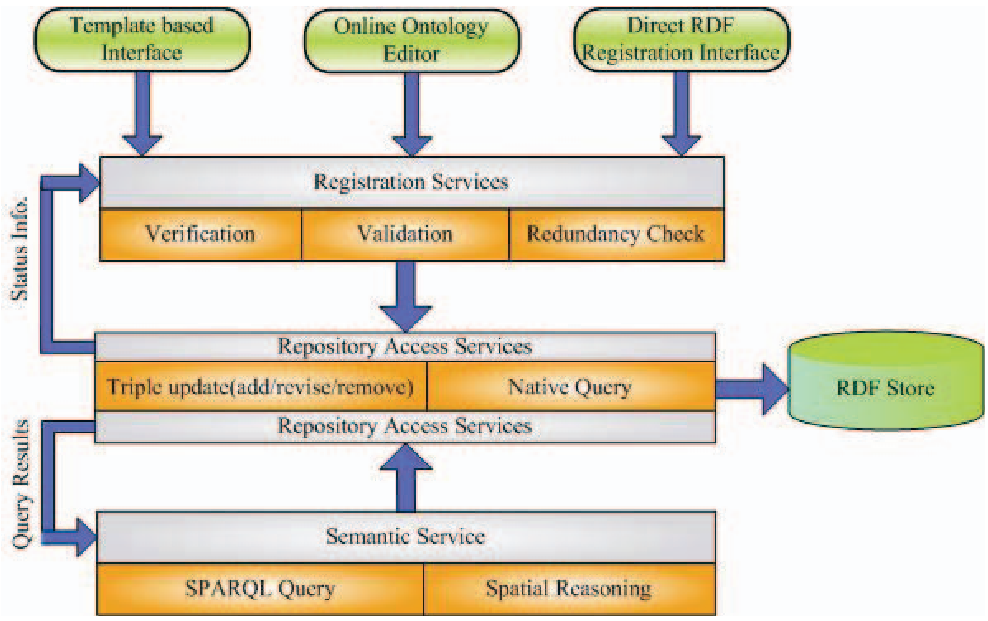

Figure 1 Architecture of the Testbed

In general, the roles of users can be divided to three types: general public, ontology engineer and ontology expert. General public do not need to understand the technical details on how knowledge base is structured. They only need to contribute by extracting semantic knowledge from unstructured web sources and input the metadata based on the template generated according to the semantic schema. Ontology engineer can operate and populate the KB through an online ontology editor. While, an ontology expert can directly register an ontology fragment in RDF format. To avoid redundancy of the backend repository, only the validated metadata will be registered to the repository by the mediation of the *registration services*. The *repository access services* are responsible for triple operations such as adding, removing or updating the KB. Note that the semantic schema of the KB is consistent with (1) ESIP datatype ontology which defines the require fields to describe a dataset, the Coordinate Reference System (CRS), distributed platform, dataset format and so on; (2) ESIP service ontology which defines the functionality, online link, input and output parameters of a spatial web service. Meanwhile, the repository access services also handle the semantic queries (instance query, subclass-superclass query or any other user defined queries) sent from *semantic services*. By the continuous contribution from researchers from the whole Earth Science community, the KB will be enriched, extended and acted as a large semantic inventory for various domain applications.

4. Prototype

Current prototype that implements the above architecture is located at: http://testbed.gmu.edu/swtestbed. We use open source framework Sesame2.3 and MySQL 5.0 as Resource Description Framework (RDF) repository; Apache Tomcat 5.5.28 and Axis 1.2 to employ services in each layer; Web Protégé as the web-based ontology editor and Java servlet and Ajax to develop online interface for semantic registration.

5. Conclusion and discussion

We have presented our urgent needs in for providing such a collaborative KB development tool for geoscientific resource registration and integration. We used semantic technologies to quickly develop and deploy an integrated registry of scientific data in the fields of ESIP datatype and service ontology. With the growing of the sophistication of use case, we will add higher level science concepts to enrich the current semantic schema. We will also make the prototype operational to support automatic service chaining and geospatial decision making.

References

ISO/IEC 11179-1:1999 Specification and standardization of data elements. Part 1: Framework for the specification and standardization of data elements (available in English only). International Standards Organisation.

R. Martell, 2009. CSW-ebRIM Registry Service - Part 1: ebRIM profile of CSW. Available online at: http://portal.opengeospatial.org/files/?artifact_id=27092 (accessed 1 November, 2009)

A. Rezgui, Z. Malik, and A. K.Sinha 2007, DIA Engine: Semantic Discovery, Integration, and Analysis of Earth Science Data, Proceedings of the Geoinformatics Conference, San Diego, CA., May 17-18, 2007.

U. Voges and K. Senkler, 2007. OpenGIS® Catalogue Services Specification 2.0.2 -ISO Metadata Application Profile. Available online at: http://portal.opengeospatial.org/files/?artifact_id=21460, OGC (accessed 1 November, 2009).

M, Zaki, R., Abdelmounaam, and S., Krishna, 2007. Semantically-enabled E-Geoscience for Knowledge Discovery. GSA Denvor Annual Meeting, Denvor, CO, 2007.