SWWG Proposal - Semantic Technologies for Integrating USGS Data

Semantic Technologies for Integrating USGS Data

Lightsom, Frances L., Varanka, Dalia E., Schweitzer, Peter N., and Gordon, Janice

Problem Statement

Semantic Web technologies represent a promising approach for integrating data from multiple USGS data systems to address interdisciplinary scientific questions. The proposed project will test and demonstrate this approach, by integrating data from five USGS data systems into an information foundation appropriate for research on aquatic habitats. A task such as this will encounter numerous challenges, particularly the integration of data with variable formats, characteristics, and meanings of data terms. Theoretical and applied solutions to resolve these problems have been proposed by an emerging community building semantic technology (Berners-Lee and others 2001).

Semantic technology is based on a data model for specifically and unambiguously describing data subjects and their relation to other entities. Specific nodes of information are programmed to link to each other according to formal semantic rules provided by an ontology (see Noy and McGuinness 2001). These automatically created networks of knowledge can access any part of their structure so that information users can query and customize the data. These functions serve to more precisely integrate data and convert information from one form to another, and thus allow a more complex context of meaning to develop around data. When connected over the Internet, these networks are often called the Semantic Web or linked data.

Objective

This proposal aims to develop and test the semantic approach to data integration by focusing on the problem of fish habitat modeling. Effective prediction of the abundance of particular species at particular locations is a primary objective of both ecology and natural resource management. Better knowledge of aquatic fish ecology and habitat requirements and improved tools for assessment and planning are needed to help conserve and rehabilitate populations throughout their native range. USGS scientists working on the National Fish Habitat Action Plan (http://www.fishhabitat.org) and aquatic aspects of the GAP Analysis Program (http://gapanalysis.usgs.gov) have these goals: (1) develop empirical species--habitat models that effectively predict the potential of specific stream reaches as habitats for important fish species, (2) describe the predicted distribution of habitats of various qualities, and (3) compare predictions with observed fish abundances. The resulting models, data, and tools will help managers assess the status of their stream habitat resources and prioritize conservation efforts. Evaluation of the model structure and predicted habitat distribution will also provide insight into the suite of conditions that best support important fish species and how those conditions vary within and between watersheds. Currently the research is conducted by discovering and collecting data, converting it to compatible formats, and using GIS systems to combine the data and create a model. We propose to investigate whether semantic techniques could automate and expedite the data discovery and integration, producing an information foundation for project scientists.

The proposed semantic demonstration project will produce an information foundation for fish habitat research that will be a “mashup” of data from multiple USGS data systems that are fragmented among the former USGS Divisions:

- The BioData Data Source contains aquatic bioassessment data (biological community and physical habitat data) using the NAWQA protocol.
- The National Land Cover Dataset provides essential information about the land adjacent to the stream reach, in the riparian buffer (locally or total upstream) and watershed (locally or total upstream). Is this land agricultural, forested, residential, or urban? Does it contain open bodies of water?
- The National Hydrography Database can be used to compute the size of the stream reach and its distance upstream from the mouth of the stream, and it also provides information about headwaters and tributaries.
- The National Elevation Dataset provides data about the slope of the stream itself as well as the topographic characteristics of its watershed.
- The Mineral Resources Database provides data about the natural geochemical characteristics of the watershed as well the impacts of mining and refining operations along the stream and in its watershed.

Methods

The proposed approach to semantic system development follows prototypes being implemented for Data.gov by researchers from Rensselaer Polytechnic Institute and Stanford University (see http://www.data.gov/semantic). The approach is iterative, with the stages diagrammed in Fig. 1.

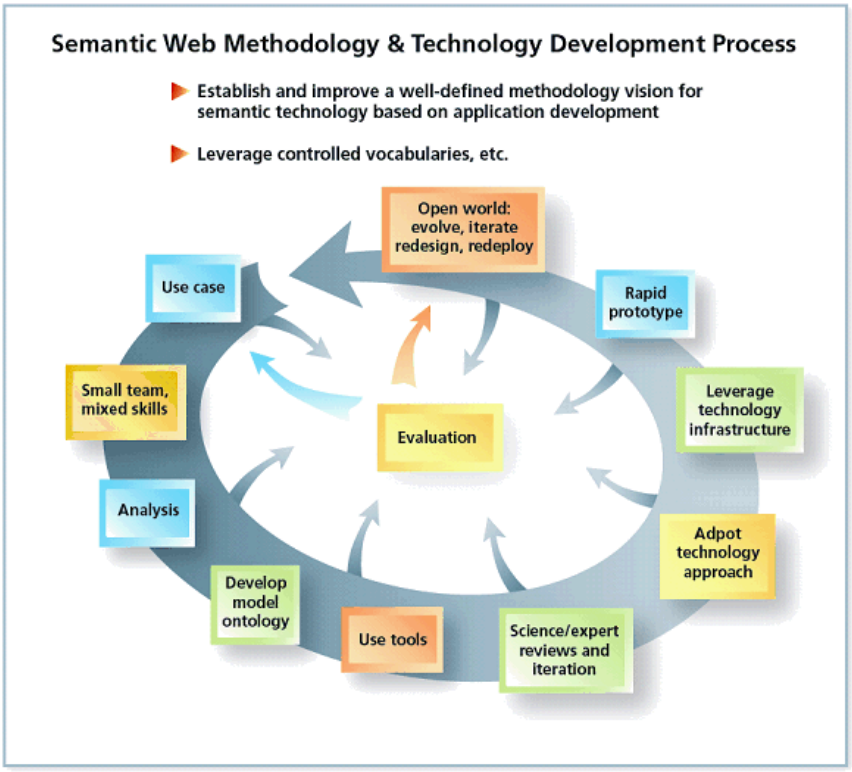

Figure 1. Semantic technology development processes developed at RPI (Peter Fox, written communication, 2011).

Stages in the prototype development

1. Develop use case(s) with the following information included
    a. Goal(s)
    b. Summary
    c. Actors
    d. Pre-Conditions
    e. Triggers
    f. Basic and Alternative Flows
    g. Post Conditions
    h. Activity Diagram
    i. Resources and Services
2. Create a project team
    a. Facilitator
    b. Domain Expert(s)
    c. Technologist(s)
    d. Data Modeler(s)
    e. Scribe
    f. Consultant(s)
3. Analysis of the use case by the project team
4. Develop information models and create ontology
    a. Logical information model
    b. Physical information model
    c. Ontology development
5. Domain expert(s) review
    a. Does the information model make sense?
6. Define technology stack to be used in the project
    a. Technical design
    b. Define system architecture requirements
7. Develop system prototype
8. Analyze and evaluate prototype system

We propose to complete one cycle of the iterative process by undertaking the following tasks:

- Refine the use case as described in Stage 1.
- Create the project team as described in Stage 2.
- Engage a consultant who is expert in the semantic development methodology, to work with us on (a) analysis of USGS data sets that will be integrated and (b) refining the focus of the use case to a manageable geographical and topical scope.
- Gather the project team for a one-week workshop hosted by CSAS in Denver to complete Stages 3 – 7 (analysis of use case, development of information models and ontology, domain experts review, definition of technology stack, design and development of working prototype). This workshop will be led by the consultant who is experienced in this method of semantic development.
- Present the working prototype to the August 2012 workshop of the USGS Community for Data Integration, for Stage 8 analysis and evaluation, as well as for sharing what we have learned with the larger group.

- Write a document that summarizes implications of this technology for USGS data systems.

Anticipated Outcomes

1. Access points for querying integrated use case data sets
2. Demonstration of semantic prototype at 2012 CDI meeting
3. Evaluation of impact on existing data systems
4. Open-file report documenting the semantic technology stack and methodology

Implications

If successful, the prototype will bring insights to the USGS science community regarding advantages and disadvantages of using semantic technology for scientific monitoring, modeling, and research.

Budget

Consulting Costs: $4,000

--Fees for leadership at a workshop and preparation.

--Travel/Mileage Costs

Project Team Costs: $12,000

--Travel Costs

--Salaries: in-kind

Hardware + System Admin Time: in-kind from CSAS

Total Costs = $16,000

References

Berners-Lee, T., Hendler, J., and Lassila, O. 2001. The Semantic Web. Scientific American, May 17, 2001, available online at http://www.scientificamerican.com/article.cfm?id=the-semantic-web

Brady, S.R., Sinha, A.K., and Gundersen, L.C., editors, 2006, Geoinformatics 2006 - Abstracts: U.S. Geological Survey Scientific Report 2006-5201, 60 p. Section 1 (p. 1-5) have a number of abstracts semantics and ontologies for geosciences.

Noy, N., and McGuinness, D.L. 2001. Ontology development 101: A guide to creating your first ontology. Stanford Knowledge Systems Laboratory Technical Report KSL-01-05 and Stanford Medical Informatics Technical Report SMI-2001-0880, available online at http://www.ksl.stanford.edu/KSL_Abstracts/KSL-01-05.html

Sinha, A. K., Malik, Z., Rezgui, A., Barnes, C.G., Lin, K., Heiken, G., Thomas, W.A., Gundersen, L.C., Raskin, R., Jackson, I., Fox, P., McGuinness, D., Seber, D., and Zimmerman, H. 2010. Geoinformatics: Transforming data to knowledge for geosciences. GSA Today, v. 20, no. 12, p. 4-10., available online at http://www.geosociety.org/gsatoday/archive/20/12/article/i1052-5173-20-12-4.htm