Post Charrette update of Semantic/Ontology Community Roadmap

Report submitted by A. Krishna Sinha
(Reviewed by Gary Berg-Cross, Anne Thessen, Nancy Wiegand, Amit Sheth, Natasha Noy, Calvin Barnes, Leo Obrst, Clinton Smyth)

Introduction

Community dialog during the EarthCube (EC) Charrette meeting (http://earthcube.ning.com/page/june-12- charrette) greatly expanded and clarified the roles of EarthCube communities to move towards common goals. Based on interaction with Charette participants, the Semantics/Ontology Community Group *(Semcg*) has updated its original roadmap to help the community converge on the primary EC goal of developing an effective data and knowledge management system.

Semantic technology is well positioned to enhance better access to data and services through the use of existing disciplinary vocabularies (ontologies), as well as by developing community endorsed new vocabularies. The success of OneGeology (http://www.onegeology.org/) is a strong reminder of the global acceptance of controlled structured vocabularies, such as GeoSciML, in developing products that have fundamentally changed the way people expect to discover and utilize resources. Based on interactions with other participants at the Charrette, we have redefined and strengthened our short and long term goals to enable a range of collaborative scientific activities that are enabled through the use and application of ontologies and the controlled vocabularies they require. We start by endorsing the need to promote sharing, access, discovery and integration of data and services through a semantic framework. This will be enabled by federating and aligning web-based access with traditional access to data center services. For scientists to enhance the scope of their research, the *Semcg* community will provide and support semantic technologies such as semantic-based search and brokering, which will enable individuals to better utilize semantic services, capabilities of the Semantic Web, and the resources that these bring to the EC community.

Our revised goals are shown in Figure 1 as a performance timeline that presents the growth of *Semcg* activities concurrent with the development of other EC activities.. These updated goals emphasize the *Semcg* objective of supporting capabilities for the EC community through the use of ontologies. In addition, these goals support collaboratively developing better semantic technologies that will aid the scientists' goals of integrating

distributed and heterogeneous data resources, including those (e.g., data level ontology) that enable 'smart search' through discovery of subsets of databases.

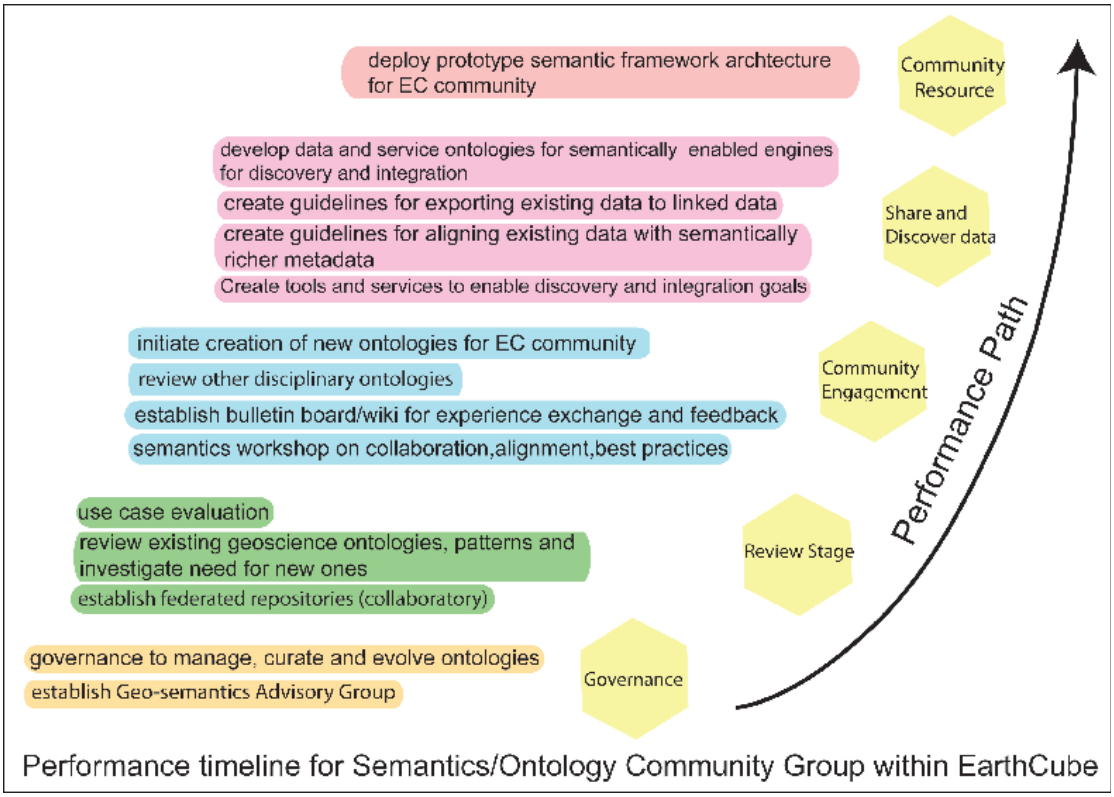

Figure 1 Performance timeline envisioned for development and deployment of semantic capabilities within the EC enterprise. Strategic growth in activities, in partnership with the broader community is recognized as essential, and dictates the timeline for deployment of semantic technologies.

Governance

We also emphasize the need to develop a governance structure early in our timeline, and suggest,

(1) The establishment of a geo-semantics advisory group that will work with the overall management architecture developed for EC. This group will provide expert advice via a consulting group of experts, and provide resources for other EC activities.

(2) The establishment of a governance structure as part of the larger EC governance framework

- to manage and curate existing ontologies and controlled vocabularies, as well as
- to evolve ontologies to meet new scientific challenges, and
- to communicate with other national and international bodies working toward these goals (such as the IUGS' Council for Geoscientific Information, responsible for GeoSciML).

These tasks constitute the need for the establishment of Governance structure within the *Semcg* community.

Review

Critical review of existing ontologies for quality and reuse is necessary prior to their application within EC activities. To support this review, an early inventory of useful ontologies will be made in the Review phase, and gaps identified relative to community needs. These gaps will be addressed during the community engagement parts of the Semcg activities. Development of a semantic collaboratory (i.e. a collaborative laboratory of scientists (geo-scientists, ontologists and computer scientists), plus a registry/repository for disciplinary ontologies with common services and tools) is suggested as a mechanism to manage ontologic resources, as well as to provide support of workshops that will enable domain scientists and ontology engineers to develop new ontologies. This collaborative environment would utilize the spectrum of use cases (based on planned Geoscience workshops) to support access to appropriate levels of ontologies required to facilitate data sharing, access and discovery. As recognized by NSF, and emphasized by EC, new funds will be available for research that is cross disciplinary, both within and external to geoscience disciplines. Therefore ontologic activities will require capabilities that go far beyond working with in-house data with sub-discipline specific vocabularies, and will require enhanced and federated repositories that will provide access to other scientific discipline ontologies. Because current metadata repositories do not provide this service, it is difficult for the EC community to conduct quality assessment prior to data usage, yet such an assessment is necessary to advance our understanding of the Earth. These ranges of activities constitute the Review stage of our performance.

Community Engagement

Once the needs of individual researchers are identified, *Semcg* will address those through focused workshops, and identify methods such as use of ontology design patterns to build flexible ontologies that are tailored to the needs of the domain scientists. Collaboration will be assisted via on-line bulletin boards and Wikis. We envision this activity to result in a collaborative environment where other EC working groups can participate and develop semantic capabilities to meet the objectives of community-developed use cases. For example, deployment of ontologies to enable semantic registration of services (semantically enabled web-services) will provide workflow communities the capability to enable and deploy automated computational tools or services linked to user-defined data. We propose to support and enhance all EC activities through development and deployment of ontologies through the use of federated ontology repositories.

We also envision working with other communities that use structured vocabulary (e.g. library, chemistry and biology communities) to enable users to discover data hosted within other disciplinary environments. These tasks will be initiated in the Community Engagement phase of our Performance path.

Share and Discover

We recognize that the infrastructure envisioned by EC is all about sharing, accessing, and discovering data within a distributed environment so that researchers can become more efficient. This efficiency will reduce the burden of spending unusually large fraction of a researcher's time in simply finding data, and thus promote more cost effective and productive use of researchers' time. ***Semcg*** will provide the semantic tagging of resources within this infrastructure to enable rapid sharing as well as discovery of data and services. Through the adoption and support of web technologies, we will encourage scientists to more easily share and discover resources on the Web. We will research and deploy technologies such as semantic brokers to enable individuals to share their data as linked open data, while maintaining control of their data. Such activities will be facilitated by ***Semcg*** developing community guidelines for aligning existing data with semantically richer metadata.

We also envision the development of data and service ontologies, as well as models for registration of data, services and data nodes, including semantically enabled engines for discovery and analysis of data. This constitutes the Share and Discover data phase of our performance plan.

Community Resource

We envision EC infrastructure as a system of systems, where tools and techniques are connected and managed within a semantic framework. In this plan, a network of databases (where data is ingested from automated acquisition technologies e.g., remote sensing instruments) is joined to a web-based linked open data environment via both syntactic and semantic connectivity. Given the ubiquitous heterogeneity of data, syntactic interoperability can occur between databases that have a common format, but semantic interoperability extends integration capabilities to heterogeneous formats by capturing the meaning of the data through the use of well defined vocabularies and data level ontologies. ***Semcg*** will support the development of such a system of systems by deploying prototypes based on complex and cross-disciplinary use case scenarios that provide the geoscience community the capability for knowledge discovery. This activity constitutes the Community Resource phase of our performance plan.