Semantic Web technology-driven Geosciences Network

Hassan A. Babaie – Department of Geosciences, Georgia State University, Atlanta, GA 30303
Email: hbabaie@gsu.edu

Summary

In this white paper, it is argued that the design and architecture of the geosciences network, as envisioned in the EarthCube framework, can be based on the Semantic Web (SW) technology. The SW technology offers the best solutions to enable the required semantic-based integration of data by providing formal (i.e., machine readable) and explicit meaning for geoscience data, thus enhancing their understanding by geoscientists. The Web Ontology Language (OWL) and its underlying RDF and RDFS languages, with the help of the Unique Resource Identifier (URI) and the HTTP protocol, allow geoscientists in different fields to build and deploy ontologies. The main purpose of building ontologies is to represent and manage domain knowledge, and to discover new knowledge, with the help of the semantic rules embedded in these languages, applying the reasoning power of the SW query languages such as SPARQL. The Semantic Web technology enables us to reach the cardinal goal of global integration of geosciences data, advancing knowledge, building effective search and query tools, and achieving interoperability and intercommunication among geoscientists and their resources at all levels in each community. We argue that the self-similarity that characterizes the research of interacting geosciences communities, and that of many geological processes, requires a fractal structuring of resources, such as software, database, ontology, and tools, over a wide range of scales, from individuals to progressively larger communities, on the proposed network.

Geosciences and the problem of interoperability

Research in geosciences is carried out by individual geoscientists or by variably sized, small to large clusters of peers that collaborate on various subjects. Clusters are assembled to improve integration of research data, and to more effectively design, develop, and deploy interoperable software, hardware, and processing and visualization tools by combining their expertise and knowledge. For stability, efficiency, and other practical reasons, such as the costs for communication, interoperability, funding, expertise, and interest, the size of these community groups, is kept to a minimum, by structuring large groups into subgroups, which are in turn split into smaller units of research, down to individual geoscientists. For the group to be stable and functioning, the subgroups, at all levels, must have ways to communicate with each other through common language, and shared, controlled vocabularies in the form of ontology (a set of knowledge-based, inter-related terms), XML markup languages, or database schema. Ontologies are needed to avoid misunderstanding of the meaning of data among the communities. The overlap, produced by the shared vocabulary, among the interacting subgroups, occurs at several orders of magnitude: within and between small sub-discipline communities (e.g., Brittle deformation in fault zones; Experimental deformation), and within and between larger fields (Structural Geology and Geophysics). The ontologies, which are built for such variably-scaled, inter-communications, may model the sub-disciplinary, disciplinary, and field knowledge, or deal with global issues for the whole geosciences community. The quality of the ontologies controls how well the data produced by these variably-sized communities are integrated.

Questions

In the context of the EarthCube framework, the main questions to be asked are as follows:

- Do we need a single, very large ontology for the geosciences community for data integration and knowledge discovery, and interoperation and intercommunication among community members, or do we need a multitude of small ontologies for each smaller community? If none of these options is good, then what is the optimum way to build the geosciences distributed, but interoperable and intercommunicating network?

- Which collective technology can achieve the optimum solution to allow us reach goals of EarthCube and GeoVision?
- Should we build a system to make the integration of smaller communities faster and easier, or should we build it to allow a wider intercommunication and interoperability among larger communities, or both?
- Should resources of a small number of nodes, on the geosciences network, handle most of the traffic by providing large number of tasks, tools, services, and ontologies, as is currently the case for example, for the Dublin Core (http://dublincore.org/) or Friend of a Friend, FOAF (http://www.foaf-project.org/) or resources should be distributed more evenly?

Fractal nature of Earth processes, geosciences research, and some geological features

The scale-invariant, self-similarity in the size and function of the geosciences research communities reflects the commonly observed fractal nature of efficient social interaction and communication (Abbott, 2001; Berner Lee and Kagal, 2008). The structure and function of the social network of geoscientists may have nodes (communities) composed of single- and multi-individual clusters of scientists, aggregated over two or more orders of magnitude. The randomly designed and built, self-similar network of inter-communicating scientists commonly exhibits a *small-world* phenomenon (Kleinberg 2000, 2006). An astonishing, characteristic feature of this kind of decentralized, small-world network is that a randomly selected node of single- or multi-individual group of scientists, connected to its neighboring nodes via short links, can, with little information, successfully communicate with far, target nodes, through a small number of intermediate nodes.

The Earth is a system composed of major, globally interconnected components (atmosphere, hydrosphere, biosphere, geosphere, cryosphere) that are in constant communication with each other through processes that involve other Earth components at progressively smaller scales, ranging from regional to atomic over several orders of magnitude (e.g., oceanic, lithospheric, subgrain). Geoscientists regularly use knowledge from other scientific fields (e.g., material science, computer science), and investigate the geophysical, geochemical, and biological processes that influence, shape, and change the major and smaller components of the Earth system. The main reason that the traditional reductionist approach – i.e., breaking down and studying the Earth system and its subsystems into their components - has served the geosciences community reasonably well, as stated in the NSF's GeoVision document, is the scale invariance and self-similarity of many of Earth's structures and processes. Self-similar processes, such as earthquake, faulting, erosion, and weathering, change the state of Earth materials, features, and structures, independent of scale. For example, structural geologists and hydrogeologists study fractures and fluid flow in fractured rocks in scales that range from microscopic to continental.

Most of the geological data, collected through field investigation, experimentation, computation, and simulation, have been published, over more than a century, in the forms of static tables and plots in scientific journals. In the past three decades, some of these data were stored in relational databases that allow local or Web-scale access and query. The geosciences legacy data include information about the behavior and properties of Earth system's multi-scaled components and characteristics of all kinds of processes that change the state of these components. Understanding of data stored in databases requires effective accessibility, query mechanism, usability, and post-search visualization by geoscientists, at all levels of their research. Generally, the integration of the heterogeneous schema and vocabulary of these distributed databases requires significant programming, often at high cost. Knowledge management systems, dependent on these distributed and heterogeneous databases are rare, and if they exist, can only be extended and scaled, with difficulty and significant cost, through constant programmatic updates.

Requirements for the geosciences network

To satisfy the EarthCube vision, the geosciences network, at the minimum, must satisfy the following requirements:

- Provide a mechanism to globally identify and integrate data acquired by variably-sized, locally-integrated but globally-distributed nodes of geoscientists who study Earth phenomena and processes at a wide range of scales, from sub-microscopic to lithospheric
- Allow integration of terms in available local and global ontologies, and their modification and extension by producing more ontologies from existing ones
- Support ontology mapping among the intercommunicating communities
- Develop a scalable, knowledge representation system to allow the overlapping geosciences community to effectively interoperate
- Support integration of globally distributed databases built by communities. Make them accessible on the Web from at least a single point of entry
- Provide ways to discover and use data stored in these local and Web-distributed databases by both geoscientists and machines (software agents)
- Develop ways to convert Web documents and paper and digital scientific publications into machine readable formats (e.g., in XML and RDF)
- Enforce inclusion of, and access to, all aspects of scientific research about the data (i.e., metadata) to accompany submitted data, for better understanding and use of the data by the community, and improving the trust in the data. These should include information provenance, experimental and computational assumptions, quality, error, precision, accuracy, uncertainty, etc., about data.
- Support distribution, discovery, use, and reuse of semantic-based knowledge models (i.e., ontologies) in all fields. Encourage the use of the controlled vocabularies, embedded in these OWL ontologies, in domain database schemas and XML markup languages. These will support knowledge representation and management, better understanding of the meaning of data, and integration of the datasets
- Make sure that data, tool, services, databases, ontologies, etc., persist, update, and remain available and functional on all parts of the network
- Provide mechanisms to visualize the data, integrate them for modeling and simulation, and provide tools and services for their further processing.
- Provide tools and services for K-12 and higher level education in geosciences

Semantic Web technology driven architecture of the network

As is noted in the NSF GeoVision document, extrapolating the meaning and significance of scientific data, acquired by studying the Earth processes and objects at small scale, to increasingly larger scales is a challenging task. Perhaps, the best way to achieve the required global access and usability is by providing mechanisms, at all scales of research that allow the meaning of data to be defined for both humans and software. Semantic Web technologies, such as RDF and OWL languages, URI, and SPARQL query language, can provide the means to satisfy the requirements for the geosciences network. The OWL and its underlying RDF and RDFS languages provide mechanisms to build ontologies, and provide reasoning capabilities for knowledge discovery. The SPARQL language allows query of the ontology-based knowledge-based systems applying the semantic rules of the ontology languages such as OWL.

Categorizing and structuring of the classes and properties, using the ontology languages, allows the domain knowledge, that is, a collection of statements that are known to be true about the real objects and their relationships in that field, to be explicitly formalized for machine processing. For example, ‘*cataclasite is a brittle fault rock*’, and ‘*fault displaces rock layer*’ are two such knowledge statements known to be true, at least by structural geologists. Each of these knowledge statements (facts) are represented in the RDF language by the ‘triple’ structure (*subject-predicate-object*), which constitutes the building block of ontologies. For example, in the second statement, ‘fault’, ‘displaces’, and ‘rock layer’ are the subject, predicate, and object, of the triple structure, respectively. Ontology is a collection of a multitude of RDF triples in which the domain and range for properties are defined. For example, in the following OWL code snippet, the voltage datatype property (attribute) is defined as the XML schema (xsd) string type for both the Cathodoluminescence and SEM domain classes. This means that both Cathodoluminescence and SEM have a property called voltage, whose value (range) can be given by an alphanumeric string. Ontology becomes a knowledge base when it is instantiated with data for its classes and properties, for example, when a value is assigned to the voltage property of the SEM and Cathodoluminescence classes below.

```
<owl:DatatypeProperty rdf:ID="voltage">
    <rdfs:domain>
      <owl:Class>
        <owl:unionOf rdf:parseType="Collection">
          <owl:Class rdf:about="#Cathodoluminescence"/>
          <owl:Class rdf:about="#SEM"/>
        </owl:unionOf>
      </owl:Class>
     </rdfs:domain>
     <rdfs:range rdf:resource="&xsd;string"/>
</owl:DatatypeProperty>
```

The Uniform Resource Identifier (URI) (the superset of the URL), used routinely in ontologies is the best way to unambiguously identify and locate any resource at local and web scales. The URI will link (i.e., dereference) to the resource with the efficient and well-tested Hypertext Transfer Protocol (HTTP). Community resources will have their own group of URI (namespace). For example, Petrology:Magma or Hydro:Aquifer will have its URI in the Petrology or Hydrogeology namespace, which maintains the vocabulary. The Hydro:Aquifer, for example, qualifies the term Aquifer to the Hydrogeology namespace, which is maintained by its corresponding Hydrogeology community. Each community will have its own domain name and resources of all kinds available on the network. Thus, building ontologies and applying other Semantic Web technologies will enable knowledge representation and management, and integration of data, and their global access and query.

Design of the user interface to access and use the resources on the geosciences network is of utmost importance. The entry to the geosciences network may start from the top level, by selecting a field (e.g., Geochemistry, Hydrogeology) under ‘Geosciences’. Upon selection of a single field, options may include the following minimum set: ‘Search’, ‘Submit’, ‘Tools’, ‘Services’, ‘Applications’, and ‘Tutorials’. The ‘Search’ option allows the user to search and query databases and knowledge bases, applying SQL and SPARQL. It also allows discovering the RDF and OWL code from ontologies. The ‘Submit’ option should allow geoscientists to submit ontologies, tools, services, etc., which become available under the ‘Search’ option. The ‘Tools’ and ‘Services’ options provide active lists (with links) to useful resources for geoscientists. Users can also submit and access software applications and tutorials to the network.

The databases, domain ontologies, computational and algorithmic tools and services, visualization and processing software, and search and query engines must operate at all levels of research. Moreover, they should be globally available and accessible, and be persistent, scalable, maintainable, updatable, and functional at all times. The network should also allow building the ‘linked data’ clouds to globally link geosciences communities in all fields, and make them available from each discipline's homepage page on the network when the field is selected. The linked data clouds and the ontologies behind them will achieve the goal of providing meaning (semantics) to data, making them understandable and useful to both machines and humans, and allowing discovery of knowledge by integrating and relating data together.

References

Abbott, A., 2001. Chaos of Disciplines. University of Chicago Press, Chicago, 259p.

Berners-Lee, T., and Kagal, L, 2008. The fractal nature of the Semantic Web. AI Magazine, Association for the Advancement of Artificial Intelligence, 23 (3), 29-34, available at: http://www.aaai.org/ojs/index.php/aimagazine/article/view/2161.

Kleinberg, J., 2000. The small-world phenomenon: An algorithmic perspective. Proc. 32nd ACM Symposium on Theory of Computing, 2000. (Also in HTML.)

Kleinberg, J., 2006. Complex Networks and Decentralized Search Algorithms. Proceedings of the International Congress of Mathematicians (ICM), 2006.