6.3. Lessons learned from semantic web prototyping in ecology

Cynthia Sims Parr[1], Joel Sachs[2], Tim Finin[2]
[1] Smithsonian Institution, [2] University of Maryland Baltimore County

We review lessons learned from the SPIRE project – Semantic Prototypes in Research Ecoinformatics (http://spire.umbc.edu) and the ongoing LepTree project (http://www.leptree.net). We include discussion of how ontologies are and could be used by the Encyclopedia of Life (EOL).

Ontologies are a means of storing and representing knowledge and have received increasing attention in recent years. OWL (the Web Ontology Language) and the related RDF (Resource Description Framework) are languages designed to specify the contexts for and logical relationships among terms. Once information in these formats is exposed on the web (the "semantic web"), specialized applications can easily find and integrate it with related information. For example, applications can determine whether "crow" in a web document refers a bird or a Native American tribe. Applications can merge data for "body mass" from different datasets using appropriate unit conversions or methodology adjustments, and they can fuse data from distributed, heterogeneous sources in response to query.

Large complex ontologies are difficult to understand and therefore difficult to re-use. Lightweight ontologies for ecological observations and invasive species enabled SPIRE to quickly prototype tools for ecoblogging and bioblitzes.

One approach to exposing and using semantic web data is to store and reason with data in traditional technologies such as relational databases and then expose it in RDF, rendering it accessible by "semantic search" crawlers (such as Swoogle) and specialized online query tools (such as TripleShop). RDF is used primarily as an interchange format. SPIRE has taken this approach, and EOL is taking this approach via the TDWG Species Profile Model standard.

Another approach is to work with data in RDF triple stores and then export it as RDF to the web. The LepTree project takes this approach. Rather than a relational data table with hundreds of cells that may be empty for any given row which may be compounded when new fields are added, RDF triples can capture rich information on an ad hoc and extensible basis. This approach is especially worthwhile in distributed environments where schemas are expected to evolve. Widespread adoption will require investment in tools that make it easy to use ontologies and export the instances. EOL is interested in using this approach for atomized descriptive data.

Currently many researchers are exploring a "linked data" vision of the semantic web. In addition to a hyperlinked web of HTML documents, data providers produce a highly interlinked web of data. One uses a Semantic Web Browser such as Disco or Tabulator rather than a standard web browser. As a user follows links, data relevant to each page are gathered for visualization and analysis. Unless data are explicitly linked however, they may not be discoverable. An advantage of this approach is that web sites can simultaneously offer both human-readable and machine-readable information, with the nature of the browser determining which version is served to a user.

Of course the best way forward depends on the desired outcomes. In general, we would recommend development of lightweight ontologies and a strategy that supports both the "semantic search" and "linked data" approaches. Data should be exposed as RDF/OWL and links to other semantic documents should be maximized.

Support is acknowledged from: U.S. National Science Foundation, MacArthur Foundation, and Sloan Foundation