DataNet Preliminary Proposal: Open Ontology Repository for Semantically Interoperable Research Communities

Kenneth Baclawski, PI, Northeastern University, Boston, MA
Mark Musen, co-PI, National Center for Biomedical Ontologies, Stanford University, Stanford, CA
Indra Neil Sarkar, co-PI, Marine Biological Laboratory, Woods Hole, MA
Bruce Bargmeyer, co-PI, Lawrence Berkeley National Laboratory, Berkeley, CA
Katherine Goodier, Senior Project Participant, NCI Information Systems, Reston, VA
Peter Yim, Senior Project Participant, CIM Engineering, Inc., San Mateo, CA
Mike Dean, Senior Project Participant, BBN Technologies, Cambridge, MA
Leo Obrst, Senior Project Participant, The MITRE Corporation, McLean, VA

Intellectual Merit. The science and engineering communities are producing very large data sets that can help solve many of the problems facing humanity and the natural environment. However, incorporating multiple data types from multiple sources to solve these problems is a significant interoperability challenge. Furthermore, documents and other artifacts created in the past can be as important as recently created data sets, but interoperability is even more difficult. The problem with such data sets is not just the differences in recording media (paper versus disks) but also the enormous changes in terminology over time. Current data sets run the risk of an even more rapid obsolescence as the meaning and formats of the data fields are forgotten or no longer available.

Semantic technologies based on logic, databases and the Semantic Web can address the problem of meaningful access to and integration of data both today and over decades and centuries. This proposal is to develop and deploy a new interoperability infrastructure called the Open Ontology Repository (**OOR**). This infrastructure will serve a vibrant community of scientific researchers with collections of controlled vocabularies and knowledge models that have been computationally encoded for data sharing in RDF, OWL, XML and other knowledge representation languages. More specifically, the open repository will support the full data management lifecycle for a virtual community of researcher groups. The collaborative Ontolog community has existed for over 6 years and continues to grow in both size and diversity. The initial metadata and data sets will support the following areas: (1) **Biomedicine**, including ontologies for genomics, proteomics, diseases, anatomy, model organisms, and other resources served by the highly successful BioPortal repository. (2) **Biodiversity**, especially the evolutionary history and taxonomic communities that strive to create dynamic syntheses of information such as the species pages in the Encyclopedia of Life. (3) **Climate and Environment**, including both natural and built environments. (4) **Human culture and sociology**.

Long-term sustainability is addressed by leveraging the vitality of a large virtual collaboration network to support a new paradigm for maintaining semantic linkages available through the Internet. Specifically, a federated knowledge repository will be deployed that can collectively correct for multiple points of failure and can foster collaborative stewardship of scientific knowledge. Particular emphasis will be given to the development of technological solutions that build on existing, proven architectures and standards. The OOR itself will be a standard as well as a federated repository, and a organization will be instituted to ensure that the OOR properly maintained on a permanent basis.

Broader Impacts. The proposed infrastructure for metadata could have a major impact on data set interoperability in general. To ensure that this potential impact is fully realized, the project will include a vigorous educational outreach program to bring other data-intensive research communities into the OOR initiative. Since the OOR will be an open, federated architecture and infrastructure, it is intended to be utilized by communities to host their own ontologies as well as to allow the communities to adapt previously established ontologies for their own purposes. Moreover, since ontologies formalize the language of a community, they can be the basis for education and training for their discipline, provided that the ontologies are properly annotated. Ontologies represent an exceptional opportunity for education and research.