Open Ontology Repository for Biology, Biodiversity, Environmental, Climatic, and Cultural Research Communities

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

The science and engineering communities are producing very large data sets that are also increasingly complex and diverse. These data sets could be the basis for solving important broadly-defined scientific problems such as understanding whole organisms, ecosystems and human populations. However, incorporating multiple data types from multiple sources to solve these problems remains a significant challenge. For example, a testable macroscopic biological hypothesis might involve the effect of environmental or climatic change on the genomic makeup of a given organism.

In addition to data sets currently being created, documents and other artifacts created in the past can have equal importance. 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 is forgotten or unavailable.

We believe in the promise of semantic technologies based on logic, databases and the Semantic Web as a means of addressing the problems of meaningful access to and integration of data both today and over decades and centuries. Such technologies enable distinguishable, computable, reusable, and sharable meaning of information artifacts, including data sets, documents and services.

We propose to develop an open ontology repository (**OOR**) of controlled vocabularies and knowledge models that have been encoded in RDF, OWL, and other knowledge representation languages. More specifically, we propose to develop an open repository for the metadata and data sets of the following communities: (1) **Biology**, especially the genomics, proteomics and other "omics" communities now served by the highly successful BioPortal repository. (2) **Biodiversity**, especially the species pages in the Encyclopedia of Life. (3) **Climate and Environment**, including both natural environments and built environments. (4) **Human culture and sociology**.

To address the issue of long-term sustainability, we propose to develop a new paradigm for maintaining semantic linkages available through the Internet. Specifically, we will develop a federated knowledge repository 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.

While these data sets provide a compelling case for the proposed OOR, the prospect of broader impacts is even more compelling. As an integral part of the proposed project, we intend to foster 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. We will pursue the exceptional opportunity for education and research represented by ontologies.