Preview of Award 0743429 - Annual Project Report

Cover

Federal Agency and Organization Element to Which Report is Submitted:	4900
Federal Grant or Other Identifying Number Assigned by Agency:	0743429
Project Title:	Semantic Enhancements for Ecological Data Management
PD/PI Name:	Matthew B Jones, Principal Investigator Shawn Bowers, Co-Principal Investigator Joshua S Madin, Co-Principal Investigator Margaret O'Brien, Co-Principal Investigator Mark P Schildhauer, Co-Principal Investigator
Submitting Official (if other than PD\PI):	N/A
Submission Date:	N/A
Recipient Organization:	University of California-Santa Barbara
Project/Grant Period:	08/01/2008 - 07/31/2014
Reporting Period:	08/01/2012 - 07/31/2013
Signature of Submitting Official (signature shall be submitted in accordance with agency specific instructions)	N/A

Accomplishments

* What are the major goals of the project?

Data for ecological and environmental studies quantify, among other things, the distribution and abundance of organisms; the processes that influence biological populations, communities, and ecosystems; and the environmental and anthropogenic drivers of these processes. Scientists increasingly rely on accessing and analyzing these diverse data collected by cross-disciplinary communities of researchers to achieve synthetic, crosscutting insights into the environment that can address issues of fundamental importance to science and society.

Despite these needs, discovering these data is difficult. The precision and recall of data searches in data repositories is not satisfactory even at current collection sizes. Data archives like the Knowledge Network for Biocomplexity (KNB), the National Biological Information Infrastructure (NBII) Metadata Clearinghouse, and the Global Change Master Directory (GCMD) rely on semi-structured metadata with fields containing largely natural-language descriptions to provide search and browsing capabilities and to allow human use and interpretation of the data. These metadata enable simple keyword searches that return results generally related to the topics of interest, but they cannot be used to perform precise searches of the data archives. Ironically data sets with more extensive (natural language) metadata are included in search results simply due to the incidental mention of a term in an ancillary part of the metadata document. These extraneous results decrease the precision of the search, seriously reducing the efficiency in researchers' finding the data they need. In addition, because natural-language metadata does not generally rely on controlled vocabularies, researchers typically classify their data sets using ad-hoc descriptive terms, reducing recall. Given the number of synonyms and overlapping terms used in scientific disciplines, searches frequently miss relevant data because the search terms do not exactly match the terms used to classify the documents.

The goals of this project are to utilize a semantic model of data and measurements in building new data management tools that can significantly improve data discovery and interpretation within ecology and environmental scieence. These tools would include tools for producing semantic metadata and attaching it to ecological data set descriptions, server software to index and reason about this semantic metadata, which in turn is used to build semantic data discovery and integration services that improve scientist's ability to locate, interpret, and repurpose scientific data from large-scale repositories such as the KNB and DataONE.

* What was accomplished under these goals (you must provide information for at least one of the 4 categories below)?

Major Activities:

During prior reporting periods, activities focused on the refinement of a core model for scientific observations (in collaboration with SONet), on the development of a prototype semantic search system for Metacat, and on a prototype semantic annotation system for Morpho.These prototypes represent a proof-of-concept for semantic search approaches and allowed us to compare multiple search strategies (see prior publications). As shown in Figure 1, we have added support to Metacat for storing and managing OWL-DL ontologies and semantic annotations, and for reasoning and search services to support different semantic-search strategies.

To implement these extensions to Metacat, we used approaches that exploit the use of formal reasoning over an ontology designed to facilitate the semantic description of scientific observations (OBOE). In our current implementation, the Jena API is used to access ontologies and ontology terms within Metacat, and Pellet is used to provide reasoning services over these ontologies (e.g., to compute class subsumption hierarchies and to ensure ontologies added to Metacat are consistent). We also extended Metcat’s XML management capabilities with support for managing semantic annotations. Ontologies and annotations added to Metacat are assigned unique identifiers (URIs), allowing both to be easily accessed through external applications (e.g., Protégé). Further, ontologies and annotations can be versioned using this URL scheme.

The Extensible Observation Ontology (OBOE) provides a high-level abstraction of scientific observations and measurements that facilitate the creation of domain-specific vocabularies for defining observation and measurement semantics. OBOE is represented using OWL-DL and enables data (or metadata) structures to be linked to domain-specific ontology concepts so that critical aspects of scientific observations can be documented—such as what kind of Entity was measured, which Characteristics of that entity were measured and by which Measurement Standards (e.g., kilograms/m^2), and what other observations provide Context for interpreting those measurements. In our approach, semantic annotations are then used to map these critical parts of a scientific observation to the data instances that are stored in a data repository (see Figure 2).

During the 2012-2013 reporting year, our major activities focused on 1) interacting with related semantics and data discovery efforts to align priorities and work, and 2) recruiting a software engineer to work on development of our semantic tools. Interactions with other semantic discovery efforts occurred through three types of activities. We continued our ongoing interactions with SONet PIs to further refine and promote the use of OBOE as a core ontology for expressing measurement semantics, Co-PIs Schildhauer and Jones played a major role in the development of semantics approaches for the DataONE project, and Schildhauer participated in EarthCube semantics working group efforts. The recruitment effort was long and difficult, as our ability to attract and retain software engineers skilled with semantic reasoning expertise is limited by the salaries we are able to offer compared to competing industry. Although we were unable to locate a new engineer for the project through multiple rounds of recruitment, we instead were able to hire replacement engineers for other projects, which has allowed us to move B.

Leinfelder back onto the Semtools effort during the current grant year (2013-2014).

Specific Objectives: During the 2012-2013 reporting period, our specific objectives were to align our initiatives with the other data discovery initiatives that have been sprouting across the environmental science community. The power of a semantic search system lies less in the particular semantic representation approach chosen, and more in the power of **shared semantics** across many different systems. Consequently, for our prototype semantic search systems to truly have an impact on the broader science community, they must be part of a shared semantics system that spans the community. Our specific objectives for the end of project are to 1) advocate for a system for shared semantics within this broader community, and 2) adapt our prototype semantic search tools (OBOE, Metacat, Morpho) to support this shared semantic system so that it can be deployed broadly within the environemntal sciences, specifically through repositories such as the KNB and DataONE.

Significant Results: During prior reporting periods, we reported on our significant published papers and software product prototype releases, including semantic search within Metacat (Figure 1, Figure 6), semantic annotation within Morpho (Figure 5), and semantic representation using the OBOE ontology (Figure 3, Figure 4). We also demonstrated the feasibility of semi-automated data integration levberaging OBOE ontologies to drive integration of syntactically heterogeneous data sets (Figure 7). See prior annual reports for details of these prototyping results.

During 2012-2013, one of our key findings is that implementing the complete semantic model of observational data that is represented by OBOE may not be practical in a production system. OBOE includes classes representing the property being measured (Characteristic), as well as the thing being measured (Entity), the standard for interpreting values (MeasurementStandard), and contextual information about the measurement (Context). Each of these requires extensions to accomodate domain-specific concepts, and these extension ontologies need to have widespread community acceptance. Also, these ontologies can grow large, with subtle differences between classes that, if interpreted or applied incorrectly in annotations, can lead to reasononing and inferencing errors. Consequently, we have found that deploying a more limited-scope ontology that just utilizes the OBOE Characteristic class has a higher probability of success in terms of community acceptance and widespread deployment in production data discovery systems.

Key outcomes or Other achievements: To date, key achievements included the products described from previous years, including the development of a semantic search system in Metacat, a semantic annotation system in Morpho, and refinement of the OBOE ontology for measurements.

* What opportunities for training and professional development has the project provided?

Through Semtools, nine students have been supported and worked on the project under the direction of Shawn Bowers at Gonzaga University, in the process gaining valuable training in computer science research: Wesley Saunders, Josie Hunter, and Jay Kudo, along with 6 other student interns during the summer of 2012.

Jay Kudo worked on ObsDB, a system for uniformly storing and querying heterogeneous observational data. Wesley Saunders has been working on a Protege plugin that simplifies the development of OBOE-compatible ontologies by providing a simple forms-based user interface for creating ontology subclasses and more complex measurement types. Josie Hunter is working on analyzing KNB data sets to determine and apply attribute similarity measures to assist in semi-automating dataset semantic annotations for datasets. This work will help efficiently provide partial annotations of existing datasets, which is a time-consuming aspect of the semantic software stack we have developed.

* How have the results been disseminated to communities of interest?

Similar to prior project years, outreach included participation in workshops and working groups held by SONet, DataONE, and Earthcube, among others, and publicaitons about the approach we have undertaken under Semtools for semantic data discovery.

* What do you plan to do during the next reporting period to accomplish the goals?

During the final reporting period of the project, we plan to fully implement and report on a production semantic data discovery system that we will deploy as part of both the KNB and DataONE data discovery systems. To accomplish this goal, we will need to create production-ready software releases for all of our major system components, and customize these to work within the context of existing systems deployed by the KNB and DataONE.

To accomplish the final phase of work this year, PIs Jones and Schildhauer and software engineer Leinfelder will work closely with researchers from SONet and DataONE who are focusing on semantics issues. This will create a collaborative team that can effectively complete the SemTools research and development work in a way that is compatible with emerging semantics initiatives and can be deployed in real-world data repositories. Towards that end, our first deliverable will focus on a final release of the OBOE ontology that allows portions of the ontology to be utilized independently of the others (primarily through relaxing cardinality constraints). Our second deliverable will focus on converting our semantic search system to utilize a SOLR index in combination with reasoners like JENA and Pellet that we utilized during the prototype phase. This will allow us to combine the large scale text-based index and search systems deployed by KNB and DataONE with the semantic reasoning capabilities exposed by JENA and Pellet. Finally, we will develop a mechanism to leverage the extensive text metadata about entities and attributes available in these repositories to streamline the process of annotating data sets against the Characteristic ontology from OBOE.

Supporting Files

Filename	Description	Uploaded By	Uploaded On
semtools-ann-report-2013-figures.pdf	Figures referenced in the report text.	Matthew Jones	12/23/2013

Products

Journals

Shawn Bowers, Joshua S. Madin, Mark P. Schildhauer (2010). Owlifier: Creating OWL-DL ontologies from simple spreadsheet-based knowledge descriptions. *Ecological Informatics*. 5 (1), 19.

Status = PUBLISHED; Acknowledgment of Federal Support = Yes ; Peer Reviewed = Yes ; DOI: doi:10.1016/j.ecoinf.2009.08.010

David Thau, Shawn Bowers, Bertram Ludäscher (2009). Merging Sets of Taxonomically Organized Data Using Concept Mappings under Uncertainty. *OTM Conferences*. 2 1103.

Status = PUBLISHED; Acknowledgment of Federal Support = Yes ; Peer Reviewed = Yes ; DOI: doi:10.1007/978-3-642-05151-7_26

Chad Berkley, Shawn Bowers, Matthew B. Jones, Joshua S. Madin, Mark Schildhauer (2009). Improving Data Discovery for Metadata Repositories through Semantic Search. *International Conference on Complex, Intelligent and Software Intensive Systems*.   1152.

Status = PUBLISHED; Acknowledgment of Federal Support = Yes ; Peer Reviewed = Yes ; DOI: doi:10.1109/CISIS.2009.122

O'Brien, M. (2010). Using the OBOE Ontology to Describe Dataset Attributes. *LTER Databits*. Fall .

Status = PUBLISHED; Acknowledgment of Federal Support = Yes ; Peer Reviewed = No ; DOI: http://databits.lternet.edu

S. Bowers, J. Kudo, H. Cao, M. Schildhauer (2010). ObsDB: A system for uniformly storing and querying heterogeneous observational data. *Proc. of the IEEE International Conference on e-Science*. 261.

Status = PUBLISHED; Acknowledgment of Federal Support = Yes ; Peer Reviewed = Yes ; DOI: doi:10.1109/eScience.2010.24

Bowers, Shawn; Cao, Huiping; Schildhauer, Mark; Jones, Matt; Leinfelder, Ben (2010). A semantic annotation framework for retrieving and analyzing observational datasets. *Proceedings of the Third Workshop on Exploiting Semantic Annotations in Information Retrieval*. 31.

Status = PUBLISHED; Acknowledgment of Federal Support = Yes ; Peer Reviewed = Yes ; DOI: doi:10.1145/1871962.1871982

Ben Leinfelder, Shawn Bowers, Margaret O’Brien, Matthew B. Jones, Mark Schildhauer (2011). Using Semantic Metadata for Discovery and Integration of Heterogeneous Ecological Data. *Proceedings of the Environmental Information Management Conference*. 92.

Status = PUBLISHED; Acknowledgment of Federal Support = Yes ; Peer Reviewed = Yes ; DOI: doi:10.5060/D2NC5Z4X

Huiping Cao, Shawn Bowers, Mark P. Schildhauer (2012). Database Support for Enabling Data-Discovery Queries over Semantically-Annotated Observational Data. *Lecture Notes in Computer Science*. 7600 198.

Status = PUBLISHED; Acknowledgment of Federal Support = Yes ; Peer Reviewed = Yes ; DOI: doi:10.1007/978-3-642-34179-3_7

Books

Book Chapters

Thesis/Dissertations

Conference Papers and Presentations

W. Saunders, S. Bowers, M. O'Brien (2011). *Protege Extensions for Scientist-Oriented Modeling of Observation and Measurement Semantics*. Proc. of the International Workshop on OWL Experiences and Directions (OWLED). San Francisco, California.

Status = PUBLISHED; Acknowledgement of Federal Support = Yes

Other Publications

Technologies or Techniques

Nothing to report.

Patents

Nothing to report.

Inventions

Nothing to report.

Licenses

Nothing to report.

Websites

Title: Semtools

URL: http://semtools.ecoinformatics.org

Description: The Semtools web site is used to describe project goals and accomplishments, disseminate this information to the broader community, and share working documents among project team members.

Other Products

Product Type: Software or Netware

Description: Jones, M., Leinfelder B., Schildhauer, M., Bowers, S., O'Brien M. 2011. Prototype semantic extensions to the Metacat data repository software system. These extensions allow Metacat servers to index and search OWL-DL ontologies and annotations referencing heterogeneous data sources to improve data discovery.

Other: Software or Netware

Product Type: Jones, M., Leinfelder B., Schildhauer, M., Bowers, S., O'Brien M. 2011. Prototype
Description: semantic extensions to the Morpho metadata edior. These extensions allow Morpho users to create and edit data set annotations that conform to OWL-DL ontologies and that can be used to improve data discovery.

Other:

Participants

What individuals have worked on the project?

Name	Most Senior Project Role	Nearest Person Month Worked
Benjamin Leinfelder	Other Professional	0
Matthew B Jones	PD/PI	1
Mark P Schildhauer	Co PD/PI	1
Margaret O'Brien	Co PD/PI	0
Joshua S Madin	Co PD/PI	0
Shawn Bowers	Co PD/PI	1

What other organizations have been involved as partners?

Name	Location
DataONE	USA
EarthCube	USA
Joint Working Group on Observational Data Semantics	USA
Juvenile Migrant Exchange Network	USA
SONet	Santa Barbara, CA

Have other collaborators or contacts been involved? Y

Impacts

What is the impact on the development of the principal discipline(s) of the project?

Through our work on Semtools, we have demonstrated improvements in the effectiveness of data discovery for large, heterogeneous data collections such as the Knowledge Network for Biocomplexity (KNB). These advances have been possible through the use of a semantic model of scientific observations (Extensible Observation Ontology) and an annotation language that is used to map relational data sources to the concepts in OBOE. The prototype system that we developed will form the basis for future work this year on a production system that will have broad applicability in the ecological and environmental sciences.

The current movement within the ecological sciences to develop ontologies for organizing and formalizing what was observed and how provides the semtools team with a good opportunity to exchange ideas about creating these ontologies. Our initial work with OBOE and the Morpho Annotation Plugin has illuminated questions about how disparate ontologies can be unified without having to conform to any one approach. As the annotation plugin matures and existing EML metadata is augmented with formal observational descriptions, locating and synthesizing heterogeneous data sources will be more efficient, and downstream analysis more accurate.

What is the impact on other disciplines?

The relative newness of knowledge representation and the use of ontologies to express and formalize information in a way that machines can ‘understand’ puts our real-life use of the technology at the forefront of the art. We are involved in the OWL API user community – a forum for both providing and soliciting support for the rapidly evolving software. Similarly, our extensive use of Protégé increases the user base directly and indirectly as we encourage collaborators to view and author domain ontologies within this application. Our work on materializing scientific data sets as large OWL graphs using conventions from the Linked Open Data community also contributes to an understanding of the scalability of linked data approaches that transcends disciplines.

What is the impact on the development of human resources?

Through collaboration with the SONet project, we have worked with SONet postdocs (Huiping Cao and Ben Adams) on the development of semantic tools for ecological data management. Cao helped develop an approach to materializing data and ontologies together in an integrated view for data subsetting. Cao has now accepted a position as a faculty member at NMSU. Ben Adams focused on the intersection between formal logic and statistical approaches to knowledge representation, and has also accepted another position.

In addition, Co-PI Bowers at Gonzaga University has worked extensively with nine different undergraduate students over the life of the project, training them in significant new skills in knowledge representation, reasoning, and software development. Student projects included work on ObsDB, a system for uniformly storing and querying heterogeneous observational data, as well as a Protege plugin that simplifies the development of OBOE-compatible ontologies by providing a simple forms-based user interface for creating ontology subclasses and more complex measurement types. Students also analyzed KNB data sets and metadata to determine and apply attribute similarity measures to assist in semi-automating dataset semantic annotations for datasets. This work has led to our conclusion that efficiently provide partial annotations of existing datasets is feasible.

What is the impact on physical resources that form infrastructure?

Nothing to report.

What is the impact on institutional resources that form infrastructure?

Development of semantic search services and deployment of those services is a critical component of institutional infrastructure. For example, many institutional repositories are being launched by University and agency libraries under the growing belief that preservation of research data is critical to scientific advances. However, as management of scientific data in institutional repositories is relatively new to most library archival teams, few have yet to realize the complexity and difficulties of managing such a heterogeneous and large information resource. The semantic tools that we have developed allow instutional reppositories to incorporate semantic annotation for research data in theit infrastructure plans, and consequently eases the burden of managing heterogeneous data. We anticipate continued growth in adoption of our semantic annotation and representation approaches for observational data as the scale and connectedness of data repositories continues to grow.

What is the impact on information resources that form infrastructure?

The project is helping to build the extensive Knowledge Network for Biocomplexity (KNB) repository, which provides thousands of data sets for use in research and educational contexts. Data from the KNB will become more accessible as the semantic search facilities that we have developed become incorporated into the production Metacat software used by the KNB. This will enable educators and researchers to more readily access KNB data and therefore facilitate science and education advances in many disciplines. The information resource represented by the KNB is the most comprehensive data archive for ecological and environemntal data in existence.

In addition, we plan to leverage our semantic search technologies to enhance discovery in the DataONE federation of data repositories. DataONE spans over 14 independent research data repositories (e.g., KNB, LTER, ORNL DAAC, etc.), and thus provides an integrated search system that exposes heterogeneous data from all of these systems. As one of NSF's major projects for data preservation, DataONE represents one of the most significant information resources for research data, and as such will benefit tremendously forom incorporation of semantic search services.

What is the impact on technology transfer?

Nothing to report.

What is the impact on society beyond science and technology?

Knowledge about science and the progress of science is critical to an effective society. Advances in the Semtools project are producing new techniques for clarifying the content and meaning of scientific observations data to make it useful for tackling cross-cutting issues that are important to society. The data that are exposed in this way become useful to many communities, including local governments and resource management agencies, non-profit organizations focused on conservation issues, and educators interested in exposing students to science approaches to societal issues.

Changes

Changes in approach and reason for change

Nothing to report.

Actual or Anticipated problems or delays and actions or plans to resolve them

Recruitment and retention for software engineers that are skilled with semantic technologies and at the available salary levels has been a continuing source of difficulty for the project. We have resolved this issue as described to NSF in our Request for No Cost Extension by moving an existing engineer (B. Leinfelder) with the requisite skills from another project to this one, and do not anticipate any further issues arising on this project.

Changes that have a significant impact on expenditures

Nothing to report.

Significant changes in use or care of human subjects

Nothing to report.

Significant changes in use or care of vertebrate animals

Nothing to report.

Significant changes in use or care of biohazards

Nothing to report.