Preview of Award 0743429 - Final Project Report

Cover |
Accomplishments |
Products |
Participants/Organizations |
Impacts |
Changes/Problems

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
Recipient Organization:	University of California-Santa Barbara
Project/Grant Period:	08/01/2008 - 07/31/2014
Reporting Period:	08/01/2013 - 07/31/2014
Submitting Official (if other than PD\PI):	N/A
Submission Date:	N/A
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 science. 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:

Activities 2008-2009. We started with refinement of our OWL-DL model for scientific observations (OBOE; Figure 2), and development of a prototype semantic search system for Metacat. This prototype was a proof-of-concept for semantic search approaches and allowed us to compare multiple search strategies. As shown in Figure 1, we 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.

In our first semantic search implementation, the Jena API was used to access ontologies and ontology terms within Metacat, and Pellet was 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.

In addition to plain-text keyword search, we implemented three different search methodologies to investigate the utility of semantic methods for scientific data discovery: (i) simple term expansion against ontologies to broaden the search terms against the metadata corpus; (ii) term expansion against semantic annotations; and (iii) structured searches that pose queries against the components of an observation described via OBOE.

Activities 2009-2010. In year 2 we created a new plug-in for the Morpho data management application that allows users to annotate data packages and also search for data that have previously been annotated. Morpho development coincided with authoring a domain-specific ontology that effectively describes data collected at the Santa Barbara Coastal LTER.

The Annotation Plugin for Morpho augments the existing data table view with a tabbed interface that highlights different aspects of the annotation process: Summary, Column, and Context. Given the complexity inherent in formally describing observational data, the team has tried to minimize confusion by keeping the focus of the annotation activity on the actual data table being annotated.

The search interface can be used to create compound nested queries that maximize search precision, and reduce false positive matches. To complement search precision, we use result ranking to maintain broad recall with the “best” or “closest” matches appearing at the beginning of the search results.

Activities 2010-2011. During 2010-2011, advances occurred in four principal areas: 1) development of a revised ontology for the Santa Barbara Coastal LTER data sets; 2) a new ‘MADLIB’ user interface approach to creating semantic data annotations; 3) advances to the semantic query plugin for the Metacat data repository system; and, 4) a new subsystem for materializing OBOE annotations as RDF graphs for use in Open Linked Data applications.

We revised the SBC marine ontology to reflect updates in the OBOE model to support the concept of ‘Measurement Types’, which are combinations of classes that bind together Entity, Characteristic, Measurement Standard, and Protocol classes to form a commonly used composite. These Measurement Types make is easy to group concepts that are repeatedly used within a project (Figure 4).

We refined the UI for semantic annotations in Morpho, particularly the user interface components for choosing ontology classes (Figure 5). For each attribute in a data set, users are asked to fill out a ‘Mad Lib’ style sentence that clarifies the Entity being measured, which characteristics of the Entity are measured, the units of the measurement, and the protocol.

We prototyped a user interface (Figure 6) for submitting semantic queries to Metacat that also includes an option to combine keyword and spatial criteria. We created a faceted search that exploits ontology subsumption hierarchies to show which queries will produce results in the browse hierarchy. Users can quickly find all of the data sets that measure a particular Characteristic, or they can find all of the data sets that contain measurements of particular Entities, or to specify data range criteria as part of the query specification (e.g., Plant Mass 'greater than 10' Grams) (Figure 7).

We experimented with fully materializing the information contained in metadata documents, data sets, ontologies, and the annotations that link them into an extended RDF graph that is compatible with the principals of the Linked Open Data approach. We found that the graphs containing instance data were sufficiently large to be impractical to query within the constraints of typical triple stores, which is the reason we used our global-as-local mediation design for the semantic search system that we constructed.

Activities 2011-2012. Activities in 2011-2012 focused on continued development, refinement, and refactoring of the OBOE model and extensions, development of ObsDB, a new web-based application for applying semantics to scientific data sets, and outreach activities about OBOE to various groups (see Outreach section). The main project activity in the first half of the project year was refactoring both OBOE and the extension ontologies for marine coastal ecology and for juvenile migrant salmon data to create shared concepts among extension ontologies for common areas such as the representation of space, time, taxa, and methods. We refactored OBOE and the extensions to allow for a modular set of classes that can be imported independently of one another and that may be usful to various extension ontologies.

ObsDB design and development. S. Bowers had 6 undergraduate student interns working on an initial release version of ObsDB, a new web-based application for applying semantics to scientific data sets. ObsDB represents an approach to create a unified repository for ontologies, data tables, and semantic annotations. The goal of this design and development work is to prototype a community-oriented site for management of observational data with explicit semantics attached.

Activities 2012-2014. In no-cost-extension in 2012-2014, we worked to move our prototype semantic annotation and search tools into production usage for data repositories, mainly the KNB Data Repository and the DataONE Federation. To accomplish this, we needed to overcome implementation barriers that made it difficult to scale semantic search up for use across the hundreds of thousands of data sets available in DataONE. First, to handle querying at scale and to integrate with the existing DataONE search system, we converted our semantic search system to utilize a SOLR index in combination with reasoners like JENA and Pellet. This allowed 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 (Figure 8).

Second, we reviewed existing standards and developed a new Open Annotation Ontology-based model to replace our original XML-based annotation syntax. We generated documentation for DataONE review at both the Core Cyber Infrastructure Team meeting in Santa Barbara, CA (June 2014) and the all-hands meeting in Albuquerque, NM (September 2014), and used these materials to drive a review of the proposed semantic search system with the science community.

Third, we developed a new MetacatUI extension for exercising semantic search facets within the existing Metacat search capabilities that utilizes the SOLR-based search API above (Figure 9), and to display the results of annotations to research scientists for review and correction (Figure 10).

Finally, we used the extensive natural language metadata about entities and attributes available in the KNB and DataONE repositories to streamline the process of annotating data sets by automatically generating semantic annotations using attribute-level metadata (top box, Figure 8). Names, labels, definitions, and units are used to query BioPortal-hosted ontologies for relevant concept matches.

Specific Objectives:

Based on the goals above, our specific objectives were to produce effective semantic tools for ecological data management and to show that semantics could be leveraged to both improve search and to enable data integration for highly heterogeneous environmental data collections. We aimed to produce several semantic tools, including:

- A refined measurement-driven ontology framework for classifying data sets to improve search
- A semantic annotation toolset that allows scientists to classify research data using that ontology
- A semantic search system that utilizes those annotations to improve precision and recall of measurement-based data searches

We accomplished each of these specific objectives, as outlined in the results and outcomes sections below.

Significant Results:

Our Metacat semantic search system added the ability to store OWL-DL ontologies and semantic annotations that link data set attributes to ontology terms and improved metadata search: (i) by expanding standard keyword searches with ontology term hierarchies; (ii) by allowing keyword searches to be applied to annotations in addition to traditional metadata; and (iii) by allowing more structured searches over annotations via ontology terms.

Figure 1 shows our semantic framework in which semantic annotaions clarify two data sets that contain similar data that is encoded differently. Metadata schemes such as the Ecological Metadata Language (EML) provide standard ways of describing the basic structural aspects of data, but not the semantics of the data set—the types of entities observed, the characteristics of these entities that were measured, and how these entities were observed in relation to each other.

A semantic annotation is a formal structure, which represents a mapping from data set values to ontology instances (i.e., individuals), and we used an XML-based syntax to represent annotation mappings like in Figure 2 showing two annotated attributes that: (i) represent observations of leaf-litter entities; (ii) measure the weight of leaf-litter (although using different weight characteristics); and (iii) use compatible but different measurement units (kilograms and grams). Figure 3 shows a more detailed example of EML attributes (bottom), semantic annotations (middle), and an OBOE ontology extension (top).

To evaluate improvements in precision and recall, we tested three search strategies.

In Keyword-Based Term Expansion, we “intercepted” keyword queries and expanded them according to the term hierarchies of stored ontologies. If a given search keyword matched a class name (i.e., as specified by the rdf:label property of the class), then the search was expanded to include the synonyms and subclasses. This form of search improved recall for documents, but also caused additional false positives due to the addition of keywords, thereby decreasing precision.

In Annotation-Enhanced Term Expansion, semantic annotations allowed individual data set attributes to be linked to ontology classes. By applying keyword searches only to annotations, search results improved precision by returning fewer false positives. Since the annotation was linked to a specific field within the metadata, data sets containing text comments in other fields were not matched, improving precision. Moreover, recall was improved due to matches expanded terms from the ontology’s class hierarchy.

In Observation-Based Structured Query, we searched for data sets via their structure -- observed entities (organism, site, etc.) and the characteristics and standards used to measure them. Queries were run using ontology classes for the observed entity, measurement characteristic, and measurement standard, providing both good recall – hitting all relevant data through appropriate use of term expansion – and good precision by exploiting the structure of OBOE annotations to find exactly the entity, characteristic, and context of interest to the user.

We found that the inherent complexity of fully describing an observational data table requires compact visualization, which we provided via a fill-in-the-blank summary of the OBOE Entity/Characteristic/Standard/Protocol of each observation. We tested annotation using marine datasets from the Santa Barbara Coastal LTER using the SBC ontology ontology that provided relevant classes. While the semantic annotation process itself produced an XML document that maps ontology classes to data table attributes, user interface development was challenging because of the complexity of the scientific concepts that we were trying to present. We found that use of ‘Mad Lib’ sentences (Figure 5) were easy to understand, succinctly presented the ontological information regarding the attribute, and could be compactly displayed when the user selected each data set attribute. In addition, each field of the MadLib dialog presented the user with a filtered view of the ontology, showing only compatible ontology terms that were relevant for that part of the annotation. Even with these advances, we found that users produced varying or inconsistent annotations for the same data files, which confounded their use in semantic search services.

Data queries (e.g., selecting all observations with diameter less than 5 cm.) were actually complex for our heterogeneous data corpus. Each of the tens of thousands of data sets had idiosyncratic schemas, and so there was no uniform relational model that could be queried. We used the OBOE ontology as a common global view against which queries were written, and the query subsystem rewrote these queries as appropriate for each local schema. The annotations thus allowed the native structure of each data set to be maintained while exposing data semantics.

This data query feature represented a powerful form of data union/integration (Figure 7). Knowing that an attribute represents a measurement of a Characteristic of a particular Entity indicates that these measurements are compatible, and therefore can be combined. In Figure 7, semantic annotations were used to drive local queries against two data sets with completely different schemas, but common semantics allowed us to produce a union data product that drew from the corresponding attributes in each of the heterogeneous data sets. This was a powerful and general data subsetting and integration approach.

We found that the general issue of having to import large ontologies to use just one or a few concepts from them is a major obstacle to ontology re-use, and one which we determined could be partially alleviated through modularization. Nevertheless, the problem of cascading imports that have far reaching implications for knowledge modeling still is a significant issue. OBOE was refactored to be comprised of a core model plus a set of extensions for Characteristics, Standards, Space, Time, and various domain Entities.

During our two no-cost extension years, our primary objectives were to transition our previously successful prototyping efforts into robust production systems at the KNB Data Repository and the DataONE federation.

To overcome scalability barriers, we re-implemented the semantic search system to use the SOLR search system in conjuntion with the previous reasoners that we had employed. By pre-processing the semantic axioms from annotations, we populated a SOLR index with only the key semantic information from the full knowlege model (e.g., by distilling annotations down to a simple concept membership index in SOLR, and making semantic search a simple lookup on a precomputed index). This allowed us to dynamically handle concept lookahead as users type search terms in the user interface (Figure 9). This dynamic interface was possible because the semantic relationships were pre-indexed. After extensive usability testing, we produced an effective semantic search UI that can easily be customized and is now incorporated in the Metacat data management application. We also found that the Open Annotation model makes an effective and extensible annotation mechanism in OWL, which makes it simple to fully materialize all the inferred axioms and incorporate these into our SOLR index.

Finally, we found that it was possible to generate useful semantic annotations by mining natural language metadata that we already have for a given data set. While these annotations are not always correct, on average we found that matching the textual metadata to the structured measurement ontologies allowed us to automatically infer measurement types for many data attributes, which in turn improves both recall and precision in measurement search as described above and overcomes the massive barrier that had been present due to the labor required to manually annotate data sets.

Key outcomes or Other achievements:

We produced 13 papers and numerous presentations on the semantics of measurement search, building upon our prior foundational work on the use of ontologies for ecological and environmental science.

We created 3 open source software tools meeting our specific project objectives. These tools are distributed from their source code repositories:

- Metacat semantic indexing and search extensions, with MetacatUI semantic search web interface
- Morpho semantic annotation extensions
- Semi-automated annotation generator

We clarified that the problem of cascading imports has far reaching implications for knowledge modeling and still is a significant issue for ontologies in general and for OBOE extension ontologies in particular. Overcoming this issue is critical to our community's ability to effectively build modular ontologies that can be re-used..

We trained nine graduate students during the course of research

The MetacatUI semantic search system has been incorporated into both the Metacat data repository system, and is being integrated into the DataONE software system. Thus, it will be in production use in major environmental data repositories in the near future as prodcution releases of these software systems ship.

* 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 worked 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 helped efficiently provide partial annotations of existing datasets, which is a time-consuming aspect of the semantic software stack we developed.

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

Outreach activities for the project have principally been through our 13 publications, and through talks at scientific conferences and workshops where we have discussed our approaches to semantically modeling scientific observations and the benefits of doing so, and common use cases. O’Brien is consulting with SBC LTER ecologists and oceanographers in the development of the domain-specific ontology. We also have distributed our software products to the community through our source code repositories for our open source products, including Metacat, MetacatUI, Morpho, and DataONE repositories.

O'Brien introduced OBOE concepts and the Santa Barbara Coastal LTER OBOE extension to the LTER Network community through several venues: the Information Managers' Committee meetings, the Network Newsletter, "Databits", and the working group tasked with developing the Network controlled vocabulary. Her activities included a demonstration of semantics tools in development and an introduction to the mapping of OBOE concepts to attribute definitions in Ecological Metadata Language (EML). She is also involved in an LTER working groups of information managers and scientists tasked with developing a controlled vocabulary for datasets. Early phases of this effort are focused on simple term taxonomies, but considering ontological concepts at this stage will greatly enhance an extension of the LTER vocabulary into a full ontology in the future.

In June 2012, M. O'Brien attended a workshop of the International Long Term Ecological Research (ILTER) Network entitled “Semantic Approaches to Discovery of Multilingual ILTER Data” at the East China Normal University in Shanghai, China. The workshop was hosted by the Chinese Ecosystem Research Network (CERN)/National Ecosystem Research Network of China (CNERN) and brought together information managers from China, Israel, UK, Korea, Taiwan, Japan, and the US. O'Brien presented the OBOE core ontology and the SBC LTER extension, and discussed mechanisms in the ontology that could be applied to international queries. A paper is in preparation.

C. Jones gave two presentations at the Pacific Northwest Aquatic Monitoring Project's data Management Leadership Team meeting on February 21, 2012. Participants included staff from the Oregon Department of Fish and Wildlife employees, USGS, Ecotrust, and Sitka Pacific Technologies (consultants to PNAMP). The first presentation was an overview of OBOE aimed at introducing the agency managers to observational ontologies. The discussion revolved around the applicability of observational ontologies in cross-agency monitoring efforts. The files are here: https://code.ecoinformatics.org/code/jmx/documents/presentations/20120221-cjones-pnamp-dmlt-oboe-salmon-overview.pdf

The second presentation was a more detailed look at the OBOE ontology itself, and the OBOE-Salmon extension ontology. This was directed at a more technical audience, and we discussed how the specific concepts are encoded as XML structures in OWL. The discussion revolved around the effort needed to integrate ontologies into the data workflow for the agencies, and how scientists in the community could and should become involved in defining the classes in the ontology. The files are here: https://code.ecoinformatics.org/code/jmx/documents/presentations/20120221-cjones-pnamp-dmlt-oboe-salmon-detail.pdf

Non-conference presentations on Semtools and SONet related work included:

O’Brien, M., Bowers, S., Jones, M., Schildhauer, M. and Leinfelder, B. 2010. SBC Extension of the OBOE Measurement Ontology. LTER Information Managers Committee Meeting, Kellogg Biological Station, Michigan State University, Sept 2010.

Jones, C., Schildhauer, M., Jones, M., O'Brien, M., Leinfelder, M., Bowers, S., Madin, J., Zimmerman, M. 2012. Using semantic technologies to help manage scientific data. Pacific Northwest Aquatic Monitoring Project Data Management Leadership Team meeting, February 21, 2012.

Jones, C., Schildhauer, M., Jones, M., O'Brien, M., Leinfelder, M., Bowers, S., Madin, J., Zimmerman, M. 2012. An observational ontology for the salmon research community. Pacific Northwest Aquatic Monitoring Project Data Management Leadership Team meeting, February 21, 2012.

Supporting Files

Filename	Description	Uploaded By	Uploaded On
semtools-ann-report-2014-figures.pdf	Supplemental Figures.	Matthew Jones	12/23/2014

Products

Books

Book Chapters

Cao, Huiping and Bowers, Shawn and Schildhauer, Mark P. (2012). Database Support for Enabling Data-Discovery Queries over Semantically-Annotated Observational Data. *Transactions on Large-Scale Data- and Knowledge-Centered Systems {VI}* 7600. Hameurlain, Abdelkader and Küng, Josef and Wagner, Roland and Liddle, Stephen W. and Schewe, Klaus-Dieter and Zhou, Xiaofang. Springer Berlin Heidelberg. Berlin, Heidelberg. 198--228. Status = PUBLISHED; Acknowledgement of Federal Support = Yes ; Peer Reviewed = Yes ; ISBN: 978-3-642-34178-6, 978-3-642-34179-3.

Thau, David and Bowers, Shawn and Ludäscher, Bertram (2009). Merging Sets of Taxonomically Organized Data

Using Concept Mappings under Uncertainty. *On the Move to Meaningful Internet Systems: {OTM} 2009* 5871. Meersman, Robert and Dillon, Tharam and Herrero, Pilar. Springer Berlin Heidelberg. Berlin, Heidelberg. 1103--1120. Status = PUBLISHED; Acknowledgement of Federal Support = Yes ; Peer Reviewed = Yes ; ISBN: 978-3-642-05150-0, 978-3-642-05151-7.

Conference Papers and Presentations

Bowers, Shawn and Cao, Huiping and Schildhauer, Mark and Jones, Matt and Leinfelder, Ben and O'Brien, Margaret (2010). *A semantic annotation framework for retrieving and analyzing observational datasets*. Proceedings of the third workshop on Exploiting semantic annotations in information retrieval - {ESAIR} '10. New York, New York, {USA}. Status = PUBLISHED; Acknowledgement of Federal Support = Yes

Thau, Dave Bowers, Shawn (2010). *Best Effort Data Exchange of Taxonomically Organized Data*. International Workshop on New Trends in Information Integration (NTII). Long Beach, California. Status = PUBLISHED; Acknowledgement of Federal Support = Yes

Jones, Matthew B. (2008). *Directions in observational data organization: from schemas to ontologies.*. Biodiversity Information Standards (TDWG) Annual Conference. Freemantle, Australia. Status = PUBLISHED; Acknowledgement of Federal Support = Yes

Schildhauer, Mark (2008). *Facilitating data interoperability within the environmental and ecological sciences through advanced semantic approaches*. Biodiversity Information Standards (TDWG) Annual Conference. Freemantle, Australia. Status = PUBLISHED; Acknowledgement of Federal Support = Yes

Berkley, Chad and Bowers, Shawn and Jones, Matthew B. and Madin, Joshua S. and Schildhauer, Mark (2009). *Improving Data Discovery in Metadata Repositories through Semantic Search*. Proceedings of {iSEEK}'09. . Status = PUBLISHED; Acknowledgement of Federal Support = Yes

Saunders, Wesley and Bowers, Shawn and O'Brien, Margaret (2011). *Protege Extensions for Scientist-Oriented Modeling of Observation and Measurement Semantics*. Proc. of the International Workshop on {OWL} Experiences and Directions. San Francisco, California. Status = PUBLISHED; Acknowledgement of Federal Support = Yes

Jones, Matthew B. (2009). *Semantic Data Integration for Heterogeneous Scientific Data*. Lifewatch WP5 Workshop on Semantic Data Integration. Amsterdam, Netherlands. Status = PUBLISHED; Acknowledgement of Federal Support = Yes

Leinfelder, Ben and Bowers, Shawn and O'Brien, Margaret and Jones, Matthew B. and Schildhauer., Mark (2011). *Using Semantic Metadata for Discovery and Integration of Heterogeneous Ecological Data*. Proceedings of the Environmental Information Management Conference ({EIM} 2011). Santa Barbara, {CA}. Status = PUBLISHED; Acknowledgement of Federal Support = Yes

Bowers, Shawn and Kudo, Jay and Cao, Huiping and Schildhauer, Mark P. (2010). *{ObsDB}: A System for Uniformly Storing and Querying Heterogeneous Observational Data*. Sixth {IEEE} International Conference on e-Science. Brisbane, {QLD}, Australia. Status = PUBLISHED; Acknowledgement of Federal Support = Yes

Inventions

Journals

Bowers, Shawn and Madin, Joshua S. and Schildhauer, Mark P. (2010). Owlifier: Creating {OWL}-{DL} ontologies from simple spreadsheet-based knowledge descriptions☆. *Ecological Informatics*. 5 (1), 19--25. Status = PUBLISHED; Acknowledgment of Federal Support = Yes ; Peer Reviewed = Yes ; DOI: 10.1016/j.ecoinf.2009.08.010

Licenses

Other Products

Software or Netware.

Jones, M., Leinfelder B., Schildhauer, M., Bowers, S., O'Brien M. 2011. Prototype 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.

Software or Netware.

Jones, M., Leinfelder B., Schildhauer, M., Bowers, S., O'Brien M. 2011. 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. The extensions are included as a core component of the open source Metacat data repository system. FOr downloads ,see https://knb.ecoinformatics.org/knb/docs/.

Other Publications

O'Brien, Margaret (2010). *Using the OBOE Ontology to Describe Dataset Attributes.*. LTER DataBits, http://databits.lternet.edu. Status = PUBLISHED; Acknowledgement of Federal Support = Yes

Patents

Technologies or Techniques

Thesis/Dissertations

Websites

Semantic Tools for Ecological Data Management
http://semtools.ecoinformatics.org

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 and with the broader science community.

Participants/Organizations

What individuals have worked on the project?

Name	Most Senior Project Role	Nearest Person Month Worked
Jones, Matthew	PD/PI	0
Bowers, Shawn	Co PD/PI	0
Madin, Joshua	Co PD/PI	0
O'Brien, Margaret	Co PD/PI	0
Schildhauer, Mark	Co PD/PI	0
Leinfelder, Benjamin	Other Professional	6

Full details of individuals who have worked on the project:

Matthew B Jones
Email: jones@nceas.ucsb.edu
Most Senior Project Role: PD/PI

Nearest Person Month Worked: 0

Contribution to the Project: As the Principal Investigator, Jones lead and organized all aspects of the project, including hiring engineers and students, overseeing subcontracts, and supervising all work. Jones worked closely with the software engineers to design and implement all tools developed in the project.

Funding Support: University of California

International Collaboration: No
International Travel: No

Shawn Bowers
Email: bowers@gonzaga.edu
Most Senior Project Role: Co PD/PI
Nearest Person Month Worked: 0

Contribution to the Project: As co-PI, Bowers participated in all aspects of project design, and was central to the development of OBOE and the initial semantic annotation system. Bowers supervised undergraduate and graduate student interns on the project.

Funding Support: This award.

International Collaboration: No
International Travel: No

Joshua S Madin
Email: madin@nceas.ucsb.edu
Most Senior Project Role: Co PD/PI
Nearest Person Month Worked: 0

Contribution to the Project: Madin was a co-PI on the project at its inception, but then moved to another University when he got a faculty position. His main role on the project was as a consultant.

Funding Support: MacQuarie University

International Collaboration: Yes, Australia
International Travel: No

Margaret O'Brien
Email: margaret.obrien@ucsb.edu
Most Senior Project Role: Co PD/PI
Nearest Person Month Worked: 0

Contribution to the Project: As co-PI, O'Brien focused on organizing and creating domain ontologies for marine science which were used in all aspects of system testing and validation, and she contributed to system design for all tools on the project.

Funding Support: None.

International Collaboration: No
International Travel: No

Mark P Schildhauer
Email: schild@nceas.ucsb.edu
Most Senior Project Role: Co PD/PI
Nearest Person Month Worked: 0

Contribution to the Project: As co-PI, Schildhauer helped to set project direction and acted as the main liaison to related ontology development projects such as SONet. Schildhauer helped design and develop the annotation models and OBOE, and contributed to design on other aspects of the project.

Funding Support: University of California

International Collaboration: No
International Travel: No

Benjamin Leinfelder
Email: leinfelder@nceas.ucsb.edu
Most Senior Project Role: Other Professional
Nearest Person Month Worked: 6

Contribution to the Project: Software development for Metacat and Morpho software products.

Funding Support: This award.

International Collaboration: No
International Travel: No

What other organizations have been involved as partners?

Name	Type of Partner Organization	Location
DataONE	Other Organizations (foreign or domestic)	USA
EarthCube	Other Organizations (foreign or domestic)	USA
Joint Working Group on Observational Data Semantics	Other Organizations (foreign or domestic)	USA
Juvenile Migrant Exchange Network	Other Organizations (foreign or domestic)	USA
SONet	Other Organizations (foreign or domestic)	Santa Barbara, CA

Full details of organizations that have been involved as partners:

DataONE

Organization Type: Other Organizations (foreign or domestic)
Organization Location: USA

Partner's Contribution to the Project:
Collaborative Research

More Detail on Partner and Contribution: DataONE Data Integration and Semantics Working Group. Co-PIs Schildhauer and O'Brien and PI Jones are collaborating with DataONE's semantics working group to develop an interoperable semantic data discovery application that focuses on ecohydrology as a use case. Collaborators from Semtools, SONet, DataONE, CUASHI, and other projects are developing a semantically integrated data query system that illustrates the power of semantics in making heterogeneous data accessible for cross-cutting science use cases. The ecohydrology use case will draw together water chemistry and biodiversity data from a variety of sources, including the CUASHI HIS system, the USGS NWIS system, the Santa Barbara Coastal LTER, and other sites.

EarthCube

Organization Type: Other Organizations (foreign or domestic)
Organization Location: USA

Partner's Contribution to the Project:
Collaborative Research

More Detail on Partner and Contribution: EarthCube Semantics and Ontologies Activity. Semtools co-PI Schildhauer has established a collaboration with geoscience semantics researchers as part of the EarthCube Semantics and Ontologies Working Group. This group is developing a vision and roadmap for the development and utility of semantics technologies within the geosciences. Schildhauer helped to co-author the EarthCube Roadmap for semantics (see publications list for citation).

Joint Working Group on Observational Data Semantics

Organization Type: Other Organizations (foreign or domestic)
Organization Location: USA

Partner's Contribution to the Project:
Collaborative Research

More Detail on Partner and Contribution: Semtools participants (Jones, Schildhauer, Bowers) have created the Joint Working Group on Observational Data Semantics with other participants from the DataONE, Data Conservancy, and SONet projects. The purpose of this Joint Working Group is to identify and pursue synergies between the projects in observational data semantics. We have held two workshops of the participants, which each resulted in a shared understanding of our varied models of observational data, as well as a joint commitment to compatible development. Future activities of the joint working group will include an emphasis on a core mode for observational data semantics, an exchange syntax for moving data and their associated semantics across systems, and demonstration prototypes of interoperability that arises from this work.

Juvenile Migrant Exchange Network

Organization Type: Other Organizations (foreign or domestic)
Organization Location: USA

Partner's Contribution to the Project:
Collaborative Research

More Detail on Partner and Contribution: The Moore Foundation has provided funding for collaboration between NCEAS and the Juvenile Salmon Migrant Exchange network (JMX), which is trying to collate and integrate salmon migration data across hundreds of research units in the Pacific Northwest. In this project, we have a half-time engineer who is developing a salmon migration ontology that can be used to describe all of the data originating from diverse research units spanning local, state, federal, tribal, academic, and non-governmental sectors. Developing this ontology has played two critical roles in the project. First, it has tremendous heuristic value in clarifying the subtle, nuanced differences in measurements being taken across projects. Second, it is being used to annotate a collection of data sets from the Washington Department of Fish and Wildlife, which in turn allows the Semtools project to demonstrate the power of semantic search and semantic data integration for these diverse institutions.

SONet

Organization Type: Other Organizations (foreign or domestic)
Organization Location: Santa Barbara, CA

Partner's Contribution to the Project:
Collaborative Research

More Detail on Partner and Contribution: We have been collaborating with the Scientific Observations Network (SONet), which is an NSF-funded INTEROP project focused on advancing a core semantic model of scientific observations. Semtools is basing much of its development effort on the observation model being evaluated in the SONet project. Initially we are using the OBOE model as the core and building an annotation framework around that model. As the SONet project progresses we hope to produce a generic solution that allows interoperability with different observation models and ontology authoring approaches. The SONet team has provided valuable insight into authoring domain-specific ontologies and defining best-practices regarding ontology construction.

Have other collaborators or contacts been involved? No

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) and DataONE, an NSF-funded DataNet partner. 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 system that we developed will form the basis for a production semantic search and annotation system that will be deployed wihtin DataONE and will have broad applicability in the ecological and environmental sciences.

The other major impact was on ontology development within the environmental sciences. Our development of the OBOE model, and our prototype work on using ontologies for data annotation, demonstrated how difficult it is to reach semantic clarity about envirnmental measurements. The effort resulted in changes to other standards such as the Observations and Measurements standard, and had a large impact on the development of the Semantic Sensor Netork (SSN) ontology that was initially developed by the W3C. The current movement within the ecological sciences to develop ontologies for organizing and formalizing what was observed and how provided 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 must be unified without having to accept the axioms from any particular model, which is still a challenge in ontology engineering.

Ultimately, as our annotation system is put into prodcution in the KNB and DataONE, we will have a large impact on the accessibility and utility of environmental data through as locating and synthesizing heterogeneous data sources will be more efficient, and downstream analysis more accurate.

What is the impact on other disciplines?

The relative rareness of real-world knowledge representation and the use of ontologies to express and formalize information in working systems 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 Semtools research and development, we have trained numerous students, postdocs, and software engineers from ecology, environmental science, and computer science about the use of semantics for environmental data managament. These graduate students and engineers will be able to incorporate these lessons on the use of semantics into all aspects of their future careers. And those that continue as academics may very well lead another generation to expand on the promise and avoid the pitfalls of semantic technologies.

What is the impact on physical resources that form infrastructure?

Although the Semtools project itself will not produce physical infrastruture per se, its software will be deployed on the KNB Data Repository and the DataONE federation data servers, thereby having a large impact on the data search facilities that are available in the US and globally.

What is the impact on institutional resources that form infrastructure?

Semtools semantic search technologies will have a major impact on the utility of environmental data repositories, and we expect it will help significantly improve the ability of non-profit operations like the KNB and DataONE to transition into useful, sustainable, and effective virtual organizations. The utility of effective data management is growing in the environmental sciences, and Semtools software will enable data repositories to meet the needs of these virtual orgnaizations. In addition, government agencies, including NSF, NOAA, NASA, and are increasingly under pressure to provide effective access to open data. For NSF in particular, data heterogeneity is a huge barrier to distribution of open data resulting from NSF research. The semantic annotation and search systems that we have produced will be able to help NSF and other agencies to meet open data mandates.

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 tens of 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. In addition, DataONE has also decided to incoporate Semtools semantic search software into the DataONE data search systems. DataONE is a major information integrator for the whole NSF DataNet program, as it has provided a common discovery service across all 4 of the active DataNet partners (Minnesota Population Center (MPC), DataNet Federation Consorium(DFC), SEAD, and DataONE), and a total of 25 disctinct data partners and networks (including other data providers such as LTER, Dryad, the ORNL DAAC repository, the Alaska Ocean Observing System, and others). As a consequence, Semtools software will have a large network effect as it gets applied to the hundreds of thousands of data sets available through the DataONE federation.

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/Problems

Changes in approach and reason for change
Nothing to report.

Actual or Anticipated problems or delays and actions or plans to resolve them
Nothing to report.

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.