Annual Report for Period: 08/2011 - 07/2012
Submitted on: 08/01/2012
Principal Investigator: Jones, Matthew B.
Award ID: 0743429
Organization: U of Cal Santa Barbara
Submitted By:
Jones, Matthew - Principal Investigator
Title:
Semantic Enhancements for Ecological Data Management

Project Participants

Senior Personnel

Name: Jones, Matthew
Worked for more than 160 Hours: Yes
Contribution to Project:
Jones is the project lead and coordinator for the project. He contributes to the design and conceptualization of software extensions to support semantic data management, and supervises development staff on the project.

Name: Schildhauer, Mark
Worked for more than 160 Hours: Yes
Contribution to Project:
Schildhauer is a co-PI on the project focusing on Knowledge Representation approaches, particularly with respect to observational data models. He is a leader in the development of OBOE, and is the principal liaison with the Scientfic Observations Network (SONet).

Name: Bowers, Shawn
Worked for more than 160 Hours: Yes
Contribution to Project:
Bowers is a co-PI on the project, helping to specify project goals especially with respect to semantic data representation and query capabilities.

Name: Madin, Joshua
Worked for more than 160 Hours: Yes
Contribution to Project:
Joshua Madin is a co-PI on the project and has helped to specify the overall semantic data management product suite. He also contributes to ontology development in the biological science area.

Name: O'Brien, Margaret
Worked for more than 160 Hours: Yes
Contribution to Project:
Margaret O'Brien is a co-PI on the project and is conceptualizing and leading the two scientific use case projects that will determine the capabilities needed by the proposed semantic data management tools.

Post-doc

Graduate Student

Undergraduate Student

Technician, Programmer

Name: Berkely, Chad
Worked for more than 160 Hours: Yes
Contribution to Project:

Berkley served as a software engineer on the project to develop the initial system prototype for semantic extensions to the Metacat system. Berkley was funded from the SEEK project but was working jointly on this project that extends SEEK concepts of semantic data query.

Name: Leinfelder, Ben

Worked for more than 160 Hours: Yes

Contribution to Project:

Ben Leinfelder has assumed the role of software engineer on the project, and is involved in designing and implementing the semantic extensions to Morpho, Metacat, and other related software subsystems.

Name: Jones, Christopher

Worked for more than 160 Hours: Yes

Contribution to Project:

Christopher Jones focused on ontology development in conjunction with Juvenile Migrant Exchange (JMX) project funded by the Moore Foundation.

Other Participant

Research Experience for Undergraduates

Organizational Partners

SONet

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.

Juvenile Migrant Exchange (JMX)

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.

DataONE

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 Semantics and Ontologies Group

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 EarchCube Roadmap for semantics (see publications list for citation).

Other Collaborators or Contacts

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.

Activities and Findings

Research and Education Activities: (See PDF version submitted by PI at the end of the report)

Activities 2008-2009

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.

Activities during the first year of the project have been focused on the refinement of a core

model for scientific observations (in collaboration with SONet), and on the development of a prototype semantic search system for Metacat. This prototype represents a proof-of-concept for semantic search approaches and will allow us to compare multiple search strategies. 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 extend 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., Prot?g?). 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).

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

After focusing on the development of a prototype semantic search system for Metacat in the first year of the project, we?ve turned our attention to the process of actually annotating observational data so that it can be discovered using these specialized semantic search strategies. Through a weekly feedback loop between the PIs and development team, we?ve created a new plug-in for the Morpho data management application that allows users to annotate packages and also to locate data packages that have previously been annotated. Morpho development has coincided with authoring a domain-specific ontology that effectively describes data collected at the Santa Barbara Coastal LTER.

At the core of our development is the Semtools API that can be used in both Metacat and Morpho for managing annotations and ontologies. The common features of the library provide an opaque interface for leveraging semantics. The abstracted and centralized API buffers us from the vagaries of rapidly evolving Knowledge Representation technologies. For instance, we were able to switch from Jena API to OWL API as the ontology management library without having to refactor other systems that were utilizing the existing Semtools

API.

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. Furthermore, we?ve adopted a fill-in-the-blank format for collecting the crucial facts that comprise a complete annotation. These self-explanatory, natural language sentences prompt the user for information and also provide built-in help.

The search interface for locating annotated data packages has been directly informed by enhancements to the fill-in-the-blank approach. Sharing similar user interface motifs makes the transition from semantic data consumer to semantic data provider all the more natural. We expect users will initially interact with the plugin?s query facilities before tackling their own data as annotators.

The search interface strikes a balance between expressivity and simplicity. Inspired by a familiar ?smart playlist? criteria building interface, the search panel 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. Having a solid interface for defining search criteria allows us to continue our focus on evaluating structured, ontology-based query strategies that exploit both explicit and inferred concept relationships.

Activities 2010-2011

During 2010-2011, project participants simultaneously advanced both our ontology development efforts and our semantic data processing tools. 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 produced a revised version of the ontology for the Santa Barbara Coastal (SBC) marine data sets that comprise one of the case studies driving Semtools work. The new ontology reflects updates in the OBOE model itself 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 for use in annotations. These Measurement Types make is easy to group concepts that are repeated used within a
project such as the SBC LTER. In addition, the new version of the SBC ontology now includes classes for a wider variety of data sets within SBC LTER, and has been refactored to separate spatial and temporal classes into their own ontologies that can be reused across projects. Figure 4 shows a selection of classes from the SBC ontology being applied in an annotation to an SBC data set. These types of annotations are serialized as XML documents that map the attributes in the SBC data set to the measured Characteristics found in the SBC ontology, and can be produced by the Morpho application and used by the Metacat semantic query plugin.

We have continued iterative development on the plugin for creating semantic annotations on data sets in Morpho. Figure 5 shows the user interface components in Morpho that we developed for choosing ontology classes. 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. This approach makes it clear how the ontology classes relate to the data attribute, and allow the user to choose each of the four OBOE facets separately.

Iterative development continued on the Metacat Semantic query plugin. We created a new user interface (Figure 6) for constructing and submitting semantic queries to Metacat that also includes an option to combine keyword and spatial criteria for a multi-faceted search as we transition to an annotated metadata corpus. Because of the detailed annotations that are available on data sets, we were able to create a faceted search that exploits ontology subsumption hierarchies to show which queries will produce results and which will not 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.

In addition, the query plugin allows users to query against the data instances as well as the annotations, which significantly increases the power of the search mechanism (Figure 7). This approach lets users specify data range criteria as part of the query specification (e.g., Plant Mass 'greater than 10' Grams), which results in a result that can find data sets that have not only particular types of measurements, but also to further hone the query to those data sets that have data values within specified ranges. This capability was created by extending the Data Manager library ? a utility developed over the years during other projects ? to query the heterogeneous databases after auto-loading them into relational tables.

We also began work on a new system that allows us to fully materialize 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. To accomplish this, we use the semantic annotations to pull information from the EML metadata documents and from the data objects themselves to produce a fully materialized OBOE model containing both the ontology classes that are relevant to a particular data set, but also the data instances themselves. This produces a very large OWL graph. This OBOE instance model can be explored like any ontology and abstracts the structural differences between different data sets that may share a similar semantic model. The materialized OBOE model can be loaded into an RDF triple store and queried using SPARQL (we did a proof of concept using AllegroGraph and a small Kelp data set). Our preliminary findings show that this approach will allow us to expose arbitrarily complex and heterogeneous data objects as part of the Linked Open Data cloud. However, we anticipate that the graphs containing instance data will be 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. We will continue to explore the use of this alternative approach while simultaneously evaluating the scalability of the Linked Data approach over the next year.

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).

Refactoring OBOE. 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. M. O'Brien and C. Jones revisited the top-level organization of OBOE, and determined that there was a need for shared concepts among extension ontologies for common areas such as the representation of space, time, taxa, and methods. In conjunction with S. Bowers, they 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. C. Jones then refactored the salmon ontology to use this new structure, and M. O'Brient regularizedthe SBC LTER extension to improve comparability between the extension and imported modules (e.g, spatial terms). See Findings for a description of the resulting refactored structure.

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. See Findings for the list of features they've implemented in ObsDB.

Findings:

Findings 2008-2009

Our semantic search system adds to Metacat the ability to store OWL-DL ontologies in addition to semantic annotations that link data set attributes to ontology terms. Our approach also extends Metacat to improve metadata search in multiple ways: (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. We describe our implementation of these extensions, and compare and contrast these different types of search for a corpus of annotated documents. As data repositories continue to grow, these tools will be instrumental in helping scientists precisely locate and then interpret data for their research needs.

Figure 1 shows the primary components of our semantic-discovery framework. The bottom of Figure 1 consists of two simple, example data sets. Although different types of data are often used within ecological analysis (e.g., raster, GIS, etc.), data sets are predominantly tabular (relational) and denote sets of related observations and measurements that were either directly collected or were the result of aggregation or analysis. Although not obvious, the example data sets in Figure 2 contain largely similar information consisting of

spatial locations divided into sub-locations (i.e., a plot or quadrat), fertilization treatment information, and weight measurements.

Metadata schemes such as the Ecological Metadata Language (EML) provide standard ways of describing implicit aspects of data sets. In Figure 2, we show a fragment of EML for describing the ?wt? and ?LL? attributes of the data sets. EML can be used to represent the basic structural aspects of data?the number of attributes, their names, and their allowable values? but 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?is either indirectly described (e.g., within the methods section of the metadata document) or are altogether missing. Metadata alone would not reveal the closely related semantics of the highlighted attributes from our sample data sets.

Semantic annotations extend EML by providing a mechanism to describe data set attributes in terms of OBOE concepts. An annotation is a formal structure, which represents a mapping from data set values to ontology instances (i.e., individuals), and an XML-based syntax is used to represent annotation mappings. As shown in Figure 2, we can see that the two annotated attributes: (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). Annotations can be used to find all data sets related to a particular concept, determine all of the concepts related to particular data set attributes, and compare data sets based on their corresponding OBOE structures (which can facilitate data integration). XML is used as an interchange format for representing annotations; in general, annotation providers will annotate data sets using higher-level metadata editors and interfaces provided through tools such as Morpho.

A more detailed example showing the various XML syntaxes used for representing EML attributes (bottom), semantic annotations (middle), and an OBOE ontology extension (top) are shown in Figure 3.

To improve overall precision and recall of Metacat searches we prototyped three new search strategies.

Keyword-Based Term Expansion. In this approach, we ?intercept? keyword queries issued to Metacat and expand them according to the term hierarchies of the stored ontologies. Specifically, if a given search keyword matches a class name (i.e., as specified by the rdf:label property of the class), then the search is expanded to include the synonyms of the class as well as the names of subclasses. This form of search alleviates the problem with simple keyword searches of not returning data sets described with synonyms or more specialized terms of the user-entered keywords. In our implementation, when a user enters a keyword search, Metacat locates synonyms and corresponding subclasses for each keyword using an ontology manager. The query is then augmented by Metacat with the expanded terms according to the given search constraints (i.e., whether terms should exactly match document terms and whether all given keywords must be present in a document) and executed against the current Metacat keyword search service.
Although this strategy improves recall for documents that may have been omitted with simple keyword searching, it can also cause additional false positives due to the addition of keywords. Thus, this approach generally increases recall, but not necessarily precision. Specifically, the set of metadata documents returned is always a superset of what is returned using the traditional Metacat keyword search.

Annotation-Enhanced Term Expansion. Semantic annotations allow individual data set attributes to be linked to one or more ontology classes. By applying keyword searches only to annotations, search results can potentially improve precision by returning fewer false positives. In annotation-enhanced search, each search term is first expanded using the ontology similar to the keyword-based term expansion. Here, when a search term matches an ontology class, we use the class and all subclasses to find matching annotations. An annotation is considered a match if it contains the corresponding classes according to the search constraints (see above). The metadata document linked to the matched annotation is returned by the search. For example, a metadata description of a data attribute described textually as ?counts of grasshoppers? would be annotated as a ?count per square meter? of ?Romalea guttata? by linking the attribute to the ontology classes that define these concepts. When the user searches for ?grasshopper?, the term is expanded to ?Romalea guttata? via the ontology?s class hierarchy, and the annotation linking the metadata attribute to the class ?Romalea guttata? becomes a match. Since the annotation is linked to a specific attribute reference within the metadata, data sets containing comments about ?grasshoppers? in other fields would not be matched. Moreover, recall is improved due to matches facilitated by descending the ontology?s class hierarchy.

Observation-Based Structured Query. Though the annotation-enhanced term expansion approach limits search to the relevant portions of metadata that describe data content (via attribute annotations), it does not take advantage of observation and measurement structures and relationships. In the observation-based query approach, users can search for data sets via their observed entities (organism, site, etc.) and the characteristics and standards used to measure them. In an observation-based search, queries are specified by

explicitly filling in an observation ?template? where ontology classes are given for the observed entity, measurement characteristic, and measurement standard. To search for data sets containing tree lengths, we would fill in an observation query template, using the tree class as the observed entity type and the length class as the measurement characteristic type. Search is performed by finding matches between the observation types of annotations and the query template, where a match includes searching subclasses of the template classes. This type of search has 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.

The search interfaces in Metacat were implemented rapidly to explore the implications of different search strategies. Our next stage of development is to design an effective user interface for composing semantic queries and then to use the semantic search engine to execute those queries.

Findings 2009-2010

In developing the Annotation Plugin for Morpho, we?ve found that the inherent complexity of fully describing an observational data table begs for a compact visualization of the annotation in-progress. Graph-based representations of the annotation ? used extensively in the SONet for illustrating OBOE ? show promise for effectively conveying the structure of

the observation model as applied to a specific data table. Moreover, the succinct fill-in-the-blank summaries of each observation and their relationship[s] to one another encourage a casual description that is in fact formally rigorous and adheres to our strict Entity/Characteristic/Standard/Protocol annotation mapping. We employ this same format in the search template used for observation-based structured searches.

Figure 1 shows a sample data set that has been partially annotated against the OBOE model. The biomass column ?wt? is annotated: ?The WetBiomass of the GiantKelp was recorded in Grams.?

Figure 2 graphs this annotation and also contextualizes the biomass observation Within a Subplot that is itself located Within a Plot. Context is an integral piece to the observation model as it determines how the data should be interpreted and analyzed.

Observation-Based Structured Query. With a corpus of well-described, fully annotated data packages, observation-based searches allow for very precise matching with a small likelihood of false positives commonly introduced by free-text descriptive fields used in keyword-only queries. By populating a template that specifies the observed Entity, the Characteristic measured, the Standard (unit) for that measurement, and the Protocol used for collecting that measurement, a query can unambiguously target data tables that contain those observations. Context relationships can be specified to further restrict the packages returned.

Domain-specific ontologies. In order to evaluate the effectiveness of the Morpho Annotation Plugin we?ve developed an extension to the OBOE ontology that can be applied to existing real-life datasets collected by the Santa Barbara Coastal LTER. O?Brien has selected a subset of data that investigates the nitrogen supply in giant kelp forests. The ontology organizes and formalizes such concepts as the observation Entity (GiantKelp), Characteristic (WetBiomass), Measurement Standard (Gram), and Protocol (wet vs. dry). Applying concepts to observations captured in tabular data columns requires a fairly comprehensive ontology. When annotating from the ontology, it is desirable to select the narrowest (most specific) concept available such that broad-concept searches can include all subclasses under that concept?s hierarchy. Axioms defined in domain ontologies can be

exploited by reasoners during both annotation authoring and annotation searching. We are poised to use Pellet for inferring indirect class subsumption hierarchies and also to validate imported ontologies as being logically consistent.

Findings 2010-2011

In developing the Morpho plugin for creating semantic annotations, we found that the semantic annotation process itself is relatively straightforward: we produce the XML document that maps ontology classes to data table attributes. However, the user interface development is challenging because of the complexity of the scientific concepts that we are trying to present in a simple user interface. The SBC ontology contains hundreds of Entities, Characteristics, Measurement Standards, and Protocols that are interrelated in a complex graph structure that is difficult to present and visualize. In addition, the user interface must show each of the four facets of the OBOE ontology (Entity, Characteristic, MeasurementStandard, Protocol) for each of the attributes in the data set. For even a simple data set with 15-20 attributes, this results in a large number of ontology classes that have to be chosen from specific subsets of OBOE extension ontologies. Thus, we developed and tested the use of the ?Mad Lib? user interface approach shown in Figure 5. We have found that these ?Mad Lib? sentences are easy to understand, succinctly present the ontological information regarding the attribute, and can be compactly displayed when the user selects each data set attribute. In addition, each field of the MadLib dialog presents the user with a filtered view of the ontology, showing only compatible ontology terms that are relevant for that part of the annotation, thereby significantly reducing the complexity of the ontology shown during annotation.

Although the data query feature of the semantic query plugin seems simple at first glance (selecting all observations with, for example, diameter less than 5 cm., it is actually a complex and demanding feat for the heterogeneous data corpus. In our system, each of the tens of thousands of data sets have their own idiosyncratic schemas, and therefore there is no uniform relational model that can be queried to select data values. Our extensions of the Data Manager Library allows us to load each of the data sets into their own tables in a relational database, but then the semantic annotations allow us to create appropriate SQL queries against the wide variety of schemas in the system. Thus, the OBOE ontology becomes a common global view against which queries are written, and the query subsystem rewrites these queries as appropriate for the local view on each data schema. The annotations allow the native structure of each data set to be maintained while exposing the semantics of the data regardless of the structure of the local schema.

This data query feature represents a simple but powerful form of data union/integration with the Data Manager library (Figure 7). Currently, knowing that an attribute represents a measurement of a Characteristic of a particular Entity provides the notion that these measurements are the same/compatible, and therefore can be combined, although at the current time we do not incorporate an evaluation of the Context of the measurement. Thus, in following Figure 7, one can see that semantic annotations can be used to drive local queries against two data sets with completely different schemas, but that the common semantics allow us to produce a union data product that draws from the corresponding attributes in each of the heterogeneous data sets. This is a powerful and general data subsetting and integration approach that can be applied to arbitrarily heterogeneous data sets as long as they have shared semantics that are expressed as annotations.

Findings 20101-2012

OBOE Refactoring. Through the work on the JMX Salmon ontology and the SBC ontology, we saw that certain extension ontologies will want to import pertinent sections of the OBOE

ontology rather than committing to the imports as all-or-nothing. For instance, both the Salmon and SBC ontologies benefit from higher-level ecological, temporal, and spatial concepts. However, the SBC ontology had more specific interest in chemistry concepts, whereas the Salmon work did not. Therefore, we re-factored the structure of OBOE to allow for more modular ontology sections that can be imported only when needed by certain extension ontologies. This 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 for ontologies in general and for OBOE extension ontologies in particular.

OBOE was refactored to be comprised of the following ontologies:
1) oboe-core.owl - provides the core OBOE modeling constructs.
2) oboe-characteristics.owl - provides general terms for characteristics that are measured either quantitatively or qualitatively, including physical characteristics, behavioral characteristics, and administrative characteristics.
3) oboe-standards.owl - provides terms for measurement standards that are common across OBOE extensions including units of measure and indices.
4) oboe-spatial.owl - provides terms relating to spatial concepts that are common across OBOE extensions. The terms are derived from the International Standards Organization (ISO) geospatial standards in the 19100 series and from the Geography Markup Language version 3.2.1.
5) oboe-temporal.owl - provides terms relating to temporal concepts that are common across OBOE extensions. The terms are derived from the International Standards Organization (ISO) geospatial standards in the 19100 series and from the Geography Markup Language version 3.2.1.

We then split out the more domain-relevant ontologies into modules as well, for import as needed:
6) oboe-chemistry - provides commonly used chemistry types in earth and environmental sciences.
7) oboe-taxa - provides taxonomic terms that are common across OBOE extensions.
8) oboe-biology - provides biological terms that are common across OBOE extensions.
9) oboe-ecology - provides ecological terms that are common across OBOE extensions.
10) oboe-environment - provides physical environment terms that are common across OBOE extensions.
11) oboe-anatomy - provides terms related to general anatomical concepts that are common across OBOE extensions. The model is derived from the Common Anatomical Reference Ontology (CARO).

Lastly, if a project truly wants all of the ontologies as a single import, we provide:
12) oboe.owl - A single ontology that imports all of the above.

The new ontology organization allows for imports of all, some, or none of the OBOE work, which is important in ontology development, since as we can see with the SWEET ontology, importing (and accepting the semantics of) all terms in an ontology can be not only cumbersome, but also perhaps incorrect for the given application. These files are all in subversion here: https://code.ecoinformatics.org/code/semtools/trunk/dev/oboe/

ObsDB development. Prototype work on ObsDB was completed by S. Bowers and interns at Gonzaga University. The ObsDB application provides the following features:
1) Allows for ontologies, tables (CSV files), and semantic annotations to be imported and stored within ObsDB
2) Supports a new annotation language, OAL, which uses a YAML-based syntax for fine-grained semantic annotations. The languages supports various constraints, and the ability

to define value-based mappings into ontology individuals and classes
3) Supports a YAML syntax for defining oboe classes and constraints
4) Provides a table (CSV file) 'shredder' that takes a CSV file, an annotation in OAL, and produces an RDF triple graph, which is stored internally using the Jena triple store technology. Uses the HermiT reasoner to classify and check consistency of the produced RDF graphs
5) Provides a specialized unit reasoner to determine when measurement units are compatible and/or convertible. The unit reasoner also applies unit conversions to measured values as well as conversions to measurement precision as needed
6) Supports a high-level query language for finding datasets that match observation/measurement types including context relations, as well as data-level conditions (e.g., tree heights greater than 10
meters). Also integrates the unit reasoner to automatically perform unit conversions (e.g., searching for datasets with Tree Diameter's in Feet, but finds datasets with convertible units, e.g., Tree Diameter's in Meters)
7) Supports 'virtual' RDF graphs, that allow observations to be selected from different graphs stored in ObsDB (via queries) to create a heterogeneous collection of observations (i.e., taken from multiple
underlying tables)
8) Supports a number of basic analytical functions to be applied to observations collections (triple graphs), including simple aggregation (avg, min, max, range, count, median). Also allows external functions in R via scripts to be applied over observation collections (graphs).

Training and Development:

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.

Outreach Activities:

Outreach activities for the project have principally been 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.

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

Presentations on Semtools and SONet related work included:

Jones, M. 2008. Directions in observational data organization: from schemas to ontologies. Biodiversity Information Standards (TDWG) Annual Conference. Freemantle, Australia. 19-25 October, 2008.

Schildhauer, M. 2008. Facilitating data interoperability within the environmental and ecological sciences through advanced semantic approaches. Biodiversity Information Standards (TDWG) Annual Conference. Freemantle, Australia. 19-25 October, 2008.

Schildhauer: Improving Data Discovery in Metadata Repositories through Semantic Search. CISIS/iSEEK. Fukuoka, Japan. March 18, 2009

Jones: Semantic Data Integration for Heterogeneous Scientific Data. Lifewatch WP5 Workshop on Semantic Data Integration. Amsterdam, Netherlands. May 18, 2009.

David Thau and Shawn Bowers. Best Effort Data Exchange of Taxonomically Organized Data. International Workshop on New Trends in Information Integration (NTII), In conjunction with ICDE, 2010.

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.

Journal Publications

Berkley C, Bowers S, Jones MB, Madin JS, Schildhauer M, "Improving Data Discovery in Metadata Repositories through Semantic Search", Proceedings of iSEEK'09. CISIS. IEEE Computer Society., p. 1152-11, vol. , (2009). Published,

Shawn Bowers, Joshua S. Madin, Mark P. Schildhauer., "Owlifier: Creating OWL-DL ontologies from simple spreadsheet-based knowledge descriptions.", Ecological Informatics, p. 19, vol. 5 (1), (2010). Published, doi:10.1016/j.ecoinf.2009.08.010

David Thau, Shawn Bowers, Bertram Lud??scher, "Merging Sets of Taxonomically Organized Data Using Concept Mappings under Uncertainty", OTM Conferences, p. 1103, vol. 2, (2009). Published,

O'Brien, M., "Using the OBOE Ontology to Describe Dataset Attributes.", LTER Databits, p. , vol. Fall 20, (2010). Published,

S. Bowers, J. Kudo, H. Cao, M. Schildhauer, "ObsDB: A system for uniformly storing and querying heterogeneous observational data", Proc. of the IEEE International Conference on e-Science. IEEE Computer Society., p. 261-268, vol. , (2010). Published,

S. Bowers, H. Cao, M. Schildhauer, M. Jones, B. Leinfelder, M. O???Brien, "A semantic annotation framework for retrieving and analyzing observational datasets.", Proc. of the Workshop on Exploiting Semantic Annotations in Information Retrieval (ESAIR). ACM Press., p. 31-32, vol. , (2010). Published,

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

B. Leinfelder, S. Bowers, M. O???Brien, M. B. Jones, M. Schildhauer, "Using Semantic Metadata for Discovery and Integration of Heterogeneous Ecological Data", Proceedings of the Environmental Information Management Conference 2011, p. , vol. , (2011). Accepted,

H. Cao, S. Bowers, M. Schildhauer, "Database Support for Enabling Data-Discovery Queries Over Semantically-Annotated Observational Data", Lecture Notes in Computer Science, p. , vol. 7600, (2012). Accepted,

Books or Other One-time Publications

EarthCube Semantics and Ontologies Working Group, "EarthCube: Roadmap for Creating the Semantic/Ontologic Infrastructure for the Geosciences", (2012). Report, Published
Editor(s): Sinha, A. Krishna
Bibliography: Report from the National Science Foundation EarthCube Semantics and Ontologies Working Group. Accessible at http://bit.ly/MkhjbF

Web/Internet Site

URL(s):
http://semtools.ecoinformatics.org
Description:

This is the main web site for the Semtools project. It is a collaborative wiki that was established to aid communication among project participants and help with organization and outreach for the project.

Other Specific Products

Contributions

Contributions within Discipline:

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.

Contributions to 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 Prot?g? 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.

Contributions to Human Resource Development:

Through collaboration with the SONet project, we have worked with the SONet postdoc (Huiping Cao) 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, and SONet will be recruiting additional postdoc for the remainder of the project.

Contributions to Resources for Research and Education:

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.

In addition, we have begun discussions with DataONE and the Data Conservancy, two major NSF projects for data preservation, about the incorporation of semantic observational data approaches into those networks. Over the next year, we expect to develop a tight collaboration with DataONE on co-development of semantic search tools that can be used in the DataONE system.

Contributions Beyond Science and Engineering:

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 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.

Conference Proceedings

Special Requirements

Special reporting requirements: None

Change in Objectives or Scope: None

Animal, Human Subjects, Biohazards: None

Categories for which nothing is reported:

Any Product

Any Conference

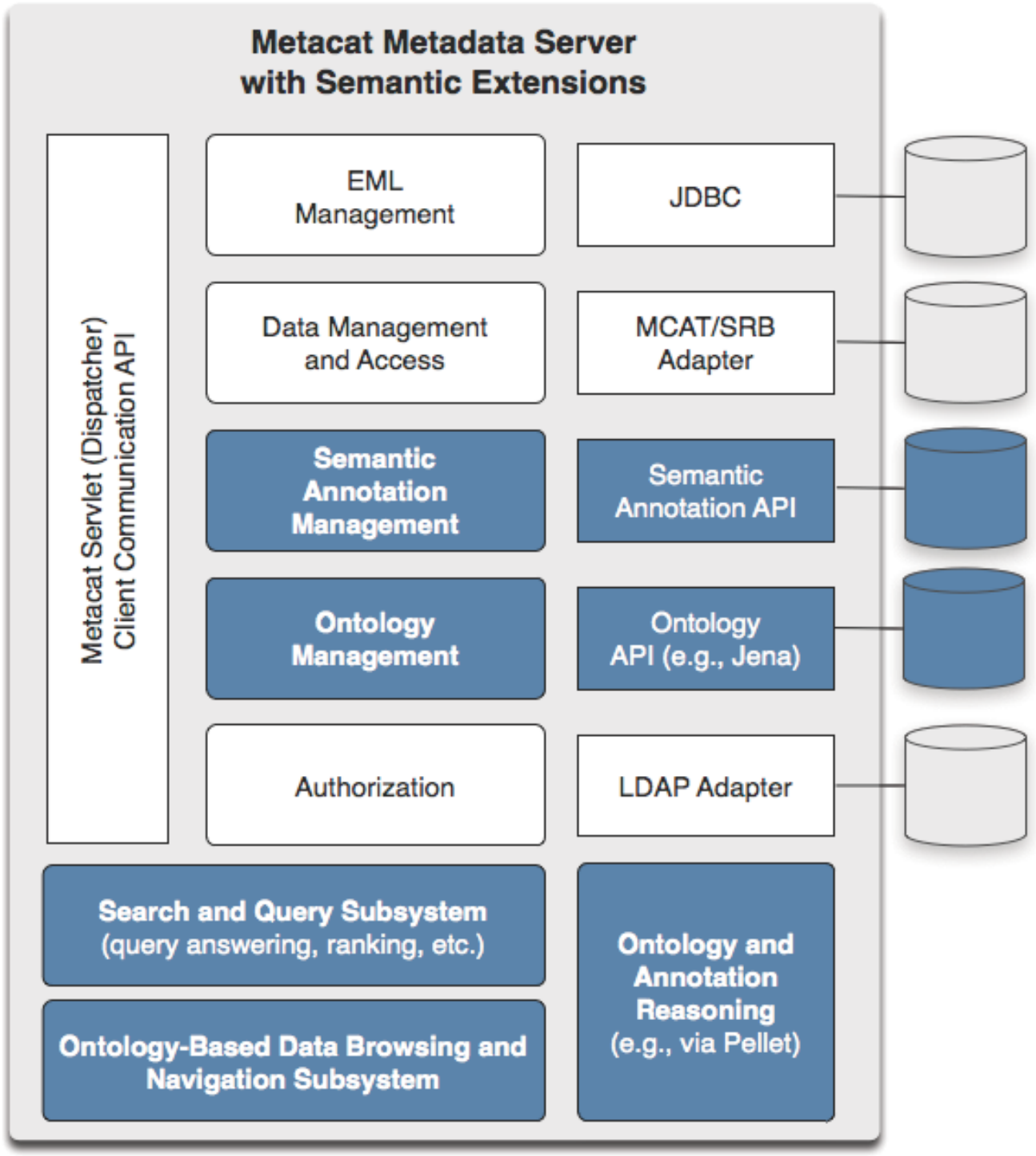


Figure 1: Semantic extensions (highlighted in blue) to the Metacat data and metadata repository support improved precision and recall in searches for scientific data sets.

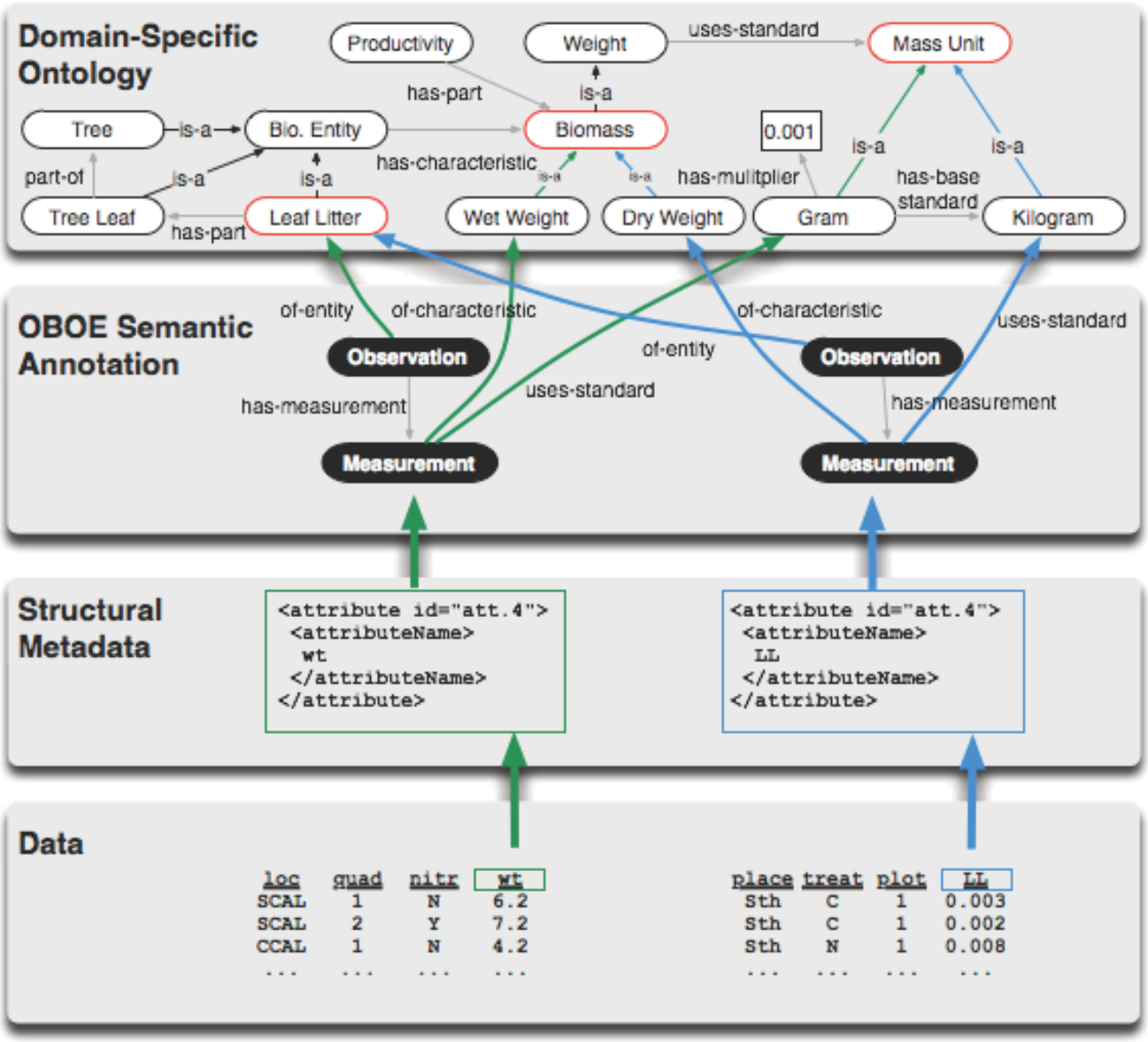


Figure 2. The components of our semantic-search framework including relational data, EML-based metadata, semantic annotations based on OBOE, and OBOE domain-ontology extensions.

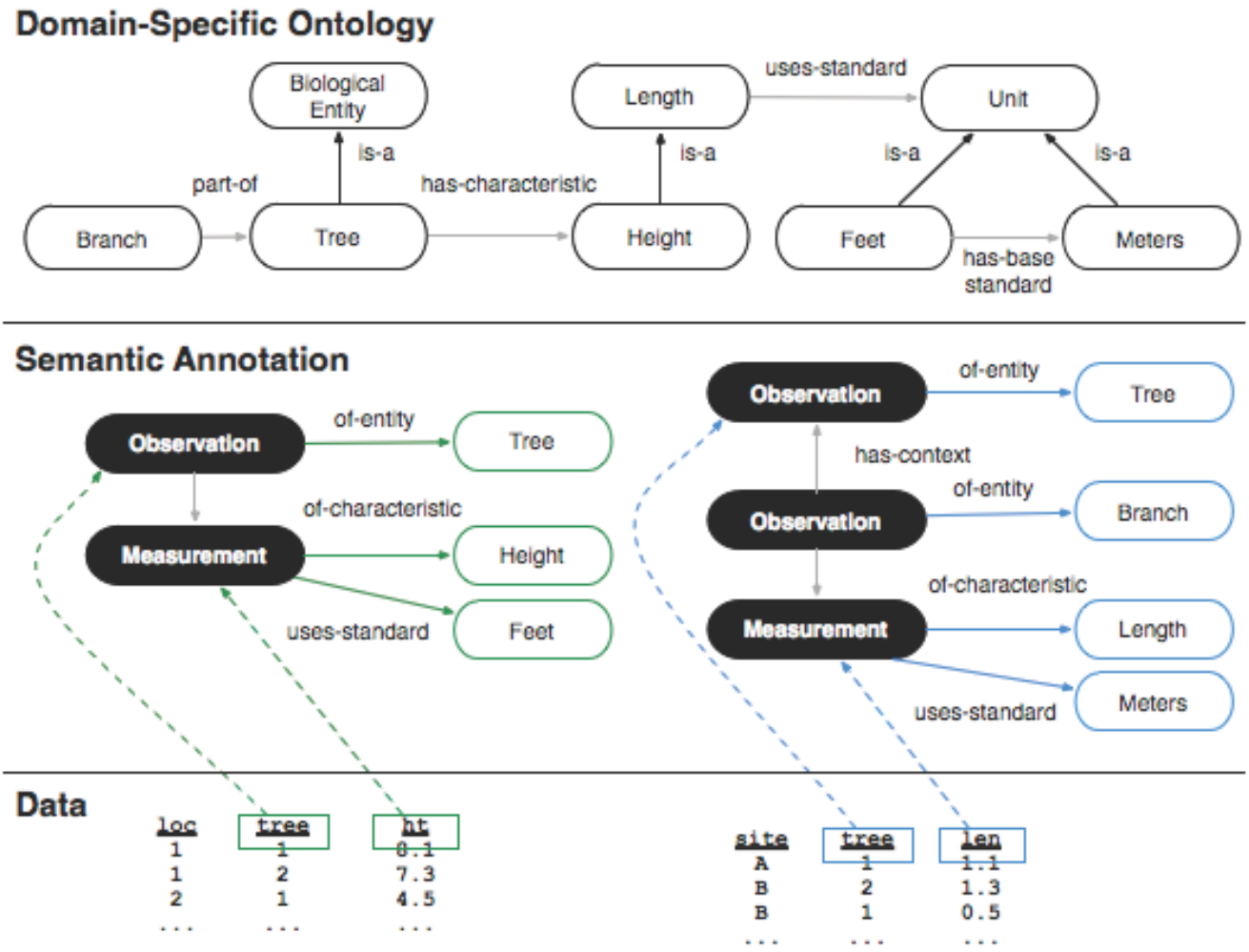
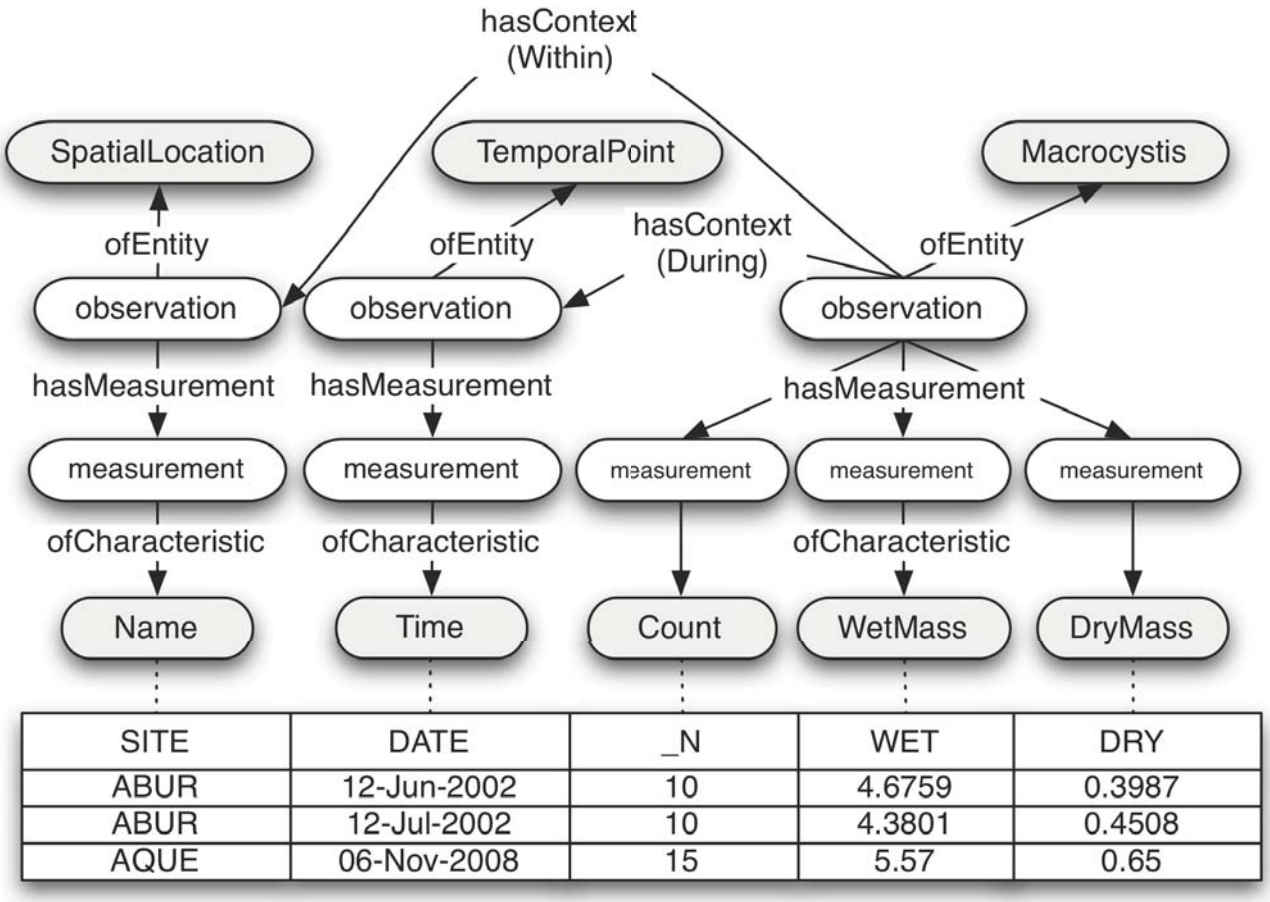


Figure 3. Example annotations demonstrating more precise search results for observation-based structured query.



SITE	DATE	_N	WET	DRY
ABUR	12-Jun-2002	10	4.6759	0.3987
ABUR	12-Jul-2002	10	4.3801	0.4508
AQUE	06-Nov-2008	15	5.57	0.65

Figure 4. Partial OBOE semantic annotation for Kelp sampling data. Shaded nodes represent ontological concepts; rectangular nodes are data table attributes mapped to OBOE measurement characteristics.

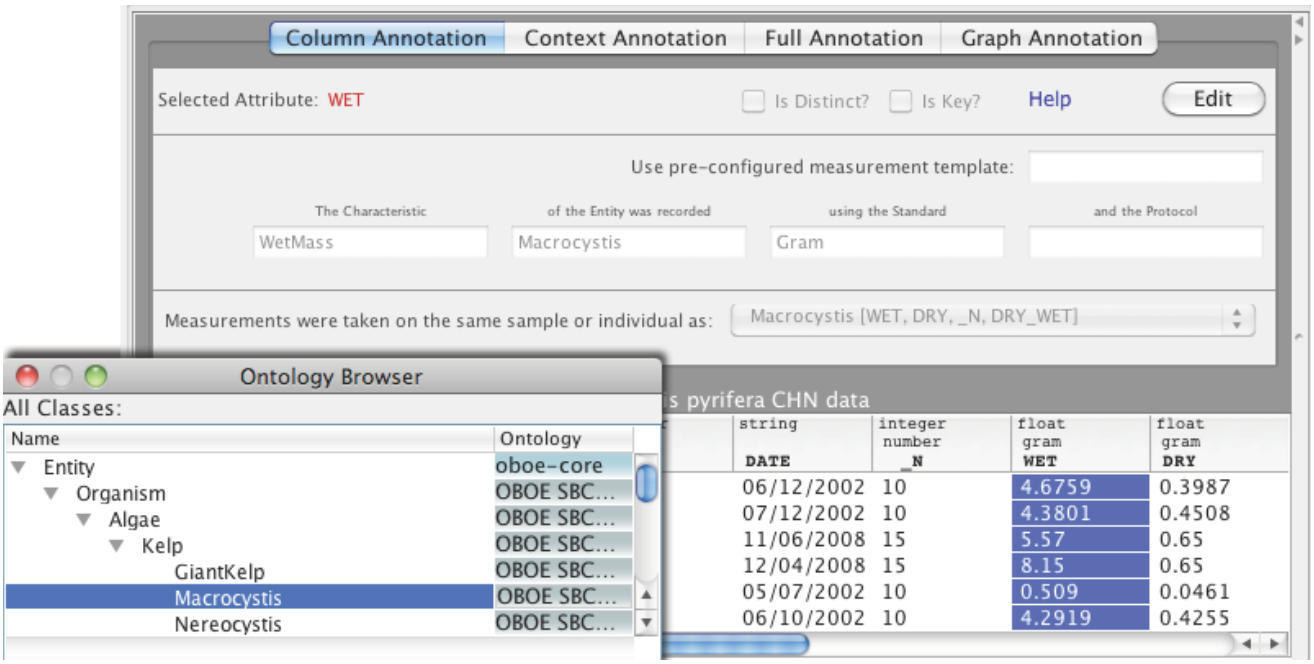


Figure 5. Morpho metadata editor with Semantic plugin. The fill-in-the-blank interface uses natural language descriptions for intuitive editing. A searchable, hierarchical browser is used to select concepts from domain-specific ontologies.

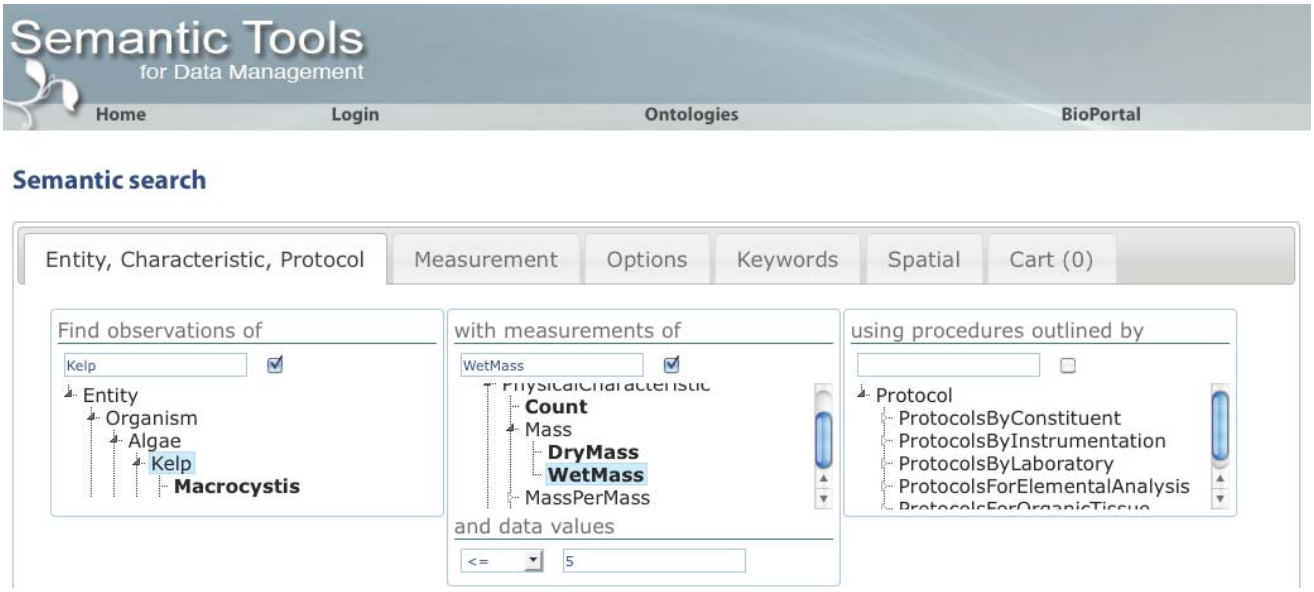


Figure 6. Semantic data query web interface. Data packages containing observations of Kelp Wet Mass less than or equal to 5 [grams] are returned. Additional search options and compound query criteria can be specified within the other tabs. Matches can be saved in the data cart for later exploration.

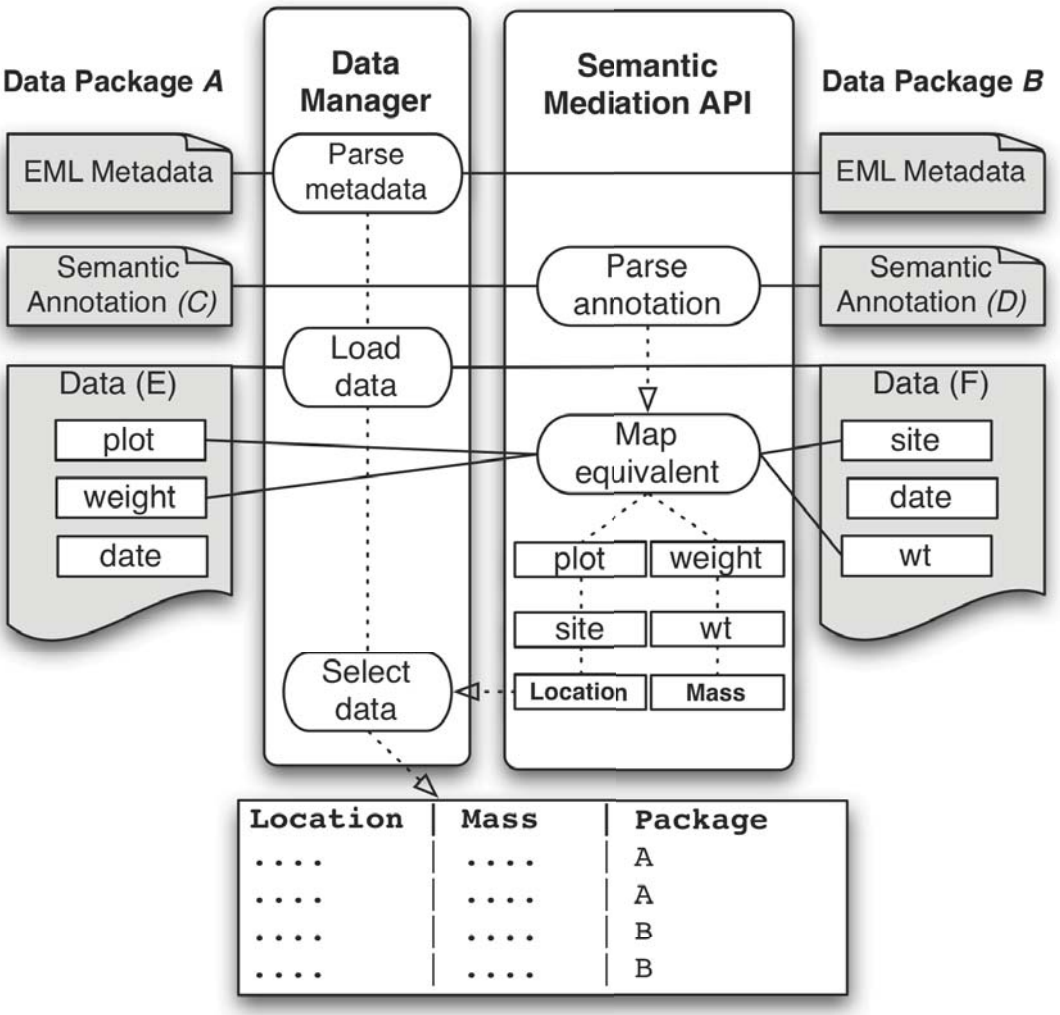


Figure 7. Integration query across multiple data packages (A, B). Annotations (C, D) determine semantically equivalent data attributes contained in the data objects (E, F). Attributes *`plot`* and *`site`* are considered equivalent measurements of the characteristic Location; attributes *`weight`* and *`wt`* both map to the same characteristic Mass. The Semantic Mediation API utilizes the Data Manager Library to load and query the source data informed by semantic similarities between the structurally disparate data attributes.