Using Semantic Metadata for Discovery and Integration of Heterogeneous Ecological Data

Ben Leinfelder[1], Shawn Bowers[2], Margaret O'Brien[3], Matthew B. Jones[1], Mark Schildhauer[1]

[1] NCEAS, University of California Santa Barbara
[2] Dept. of Computer Science, Gonzaga University
[3] Marine Science Institute, University of California Santa Barbara
{leinfelder, jones, schild}@nceas.ucsb.edu, bowers@gonzaga.edu, mob@msi.ucsb.edu

***Abstract*—Effective discovery and integration of ecological data within data management systems requires rich semantic information that can describe and relate the types of information contained within disparate data sets. Within the Semtools project, we have developed approaches for expressing and representing semantic annotations of data sets for supplementing attribute and data-level metadata with terms drawn from domain-specific ontologies. Annotations provide a formal mechanism that can be used together with reasoning systems to enhance existing data discovery and integration approaches. We describe extensions to the Ecological Metadata Language (EML) and associated tools for storing and using semantic annotations. Specifically, we describe new user interface components implemented within the Morpho metadata editor for capturing user-supplied semantic annotations, extensions to the Metacat system for storing and accessing annotations and corresponding OWL-DL ontologies, and a new API within Metacat that uses annotation metadata to provide concept-based search and integration of data sets.**

***Keywords*—ontologies; annotation; data discovery and integration**

I. Introduction

A major challenge in environmental information management concerns providing effective approaches for the discovery and integration of heterogeneous data sets. For instance, locating and combining relevant observational data are often critical and time-consuming steps for researchers studying phenomena at broad spatial, temporal, and biological scales [1], [2]. The underlying data sets used within such studies frequently differ in subtle and complex ways, due in part to the protocols used for data collection, the types of observations made, and the experimental and other contextual information associated with the data set. These differences in turn can lead to structural and semantic heterogeneity among data sets that make them hard to discover using current data management approaches and require considerable manual effort by researchers needing to combine data sets.

A number of recent efforts within the earth and environmental informatics communities are adopting the notion of an *observation* as a key modeling concept for enabling improved discovery and integration of scientific data [3]–[7]. These approaches provide higher-level observational data models for describing and representing observations and measurements found in underlying data sets by defining common "core" concepts such as the entities or features being observed, measurement units and protocols, and context relationships between observations [3], [7]. A major goal of these approaches is to enable interoperability and uniform access to data by abstracting away the underlying representation details that often impede integration across scientific data sets.

In this paper we describe extensions to the Ecological Metadata Language (EML) [8] and supporting tools for enabling improved discovery and integration of ecological data sets. Our work is based on the Extensible Observations Ontology (OBOE) [7], [9], which represents a generic observational model implemented in OWL-DL [10] for describing domain-specific observation and measurement types. Our approach adds additional metadata in the form of semantic annotations that link attributes within data sets to OBOE terms for describing the implicit observation and measurement types found within data sets. Semantic annotations are executable in the sense that they can be used to convert a data set into a collection of observation and measurement instances, providing a more uniform representation for expressing queries and performing integration. To support the creation of annotations, we have extended the Morpho metadata editor [11] with a high-level user interface as well as the Metacat data catalog [12] for storing and querying annotations through a new Semantic Mediation API. This API can also be used to perform basic data-level integration tasks using our prior work on the EML Data Manager Library [13].

The rest of this paper is organized as follows. Sec. II briefly describes the various components used within our approach including the extensions we have developed for Morpho and Metacat to support semantic annotation. Sec. III describes the types of data discovery queries and integration services supported by our framework. Sec. IV briefly describes related work, and we summarize our contributions in Sec. V.

II. Semtools Framework

The Semtools project has focused development efforts on three main components: a Java library for accessing and manipulating OBOE ontology extensions and semantic annotations, an annotation plugin for the Morpho metadata editor, and query extensions for the Metacat data catalog. Below we

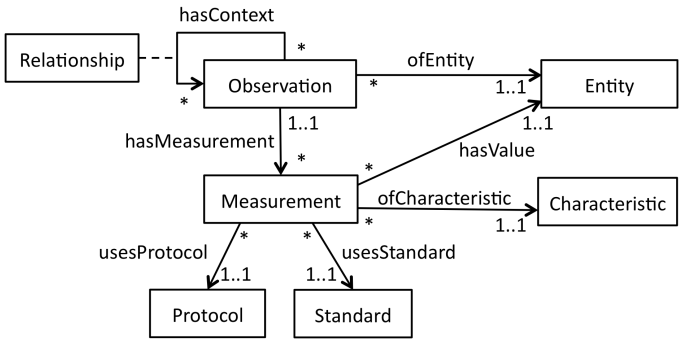

Fig. 1. Main classes and properties of the extensible observation ontology (OBOE). While shown using the Unified Modeling Language (UML), the model is defined as an OWL-DL ontology.

describe these components and give a brief overview of the OBOE model as well as the semantic annotation approach used in Semtools. For a more in-depth presentation of OBOE see [7], [9].

A. The OBOE Observational Model

Fig. 1 shows the main modeling constructs of OBOE (see: http://ecoinformatics.org/oboe/oboe.1.0/oboe-core.owl). An *observation* is made of an *entity* (e.g., biological organisms, geographic locations, environmental features) and serves to group a set of measurements together to form a single "*observation event*". A *measurement* assigns a value to a *characteristic* of the observed entity (e.g., the weight of a plant), and can also include *standards* (e.g., units as well as standards for coded values) and collection *protocols*. An observation can occur within the surrounding *context* of other observations (e.g., as part of a temporal or spatial context), and context may include a named relationship (e.g., "partOf", "within") that existed during the observation event. A key feature of OBOE is that it allows properties (characteristics and relationships) of entities to be asserted without being interpreted as *inherently* (i.e., always) true of the entity. Depending on the context in which the entity was observed or how the measurements were performed, an entity's properties may take on different values. OBOE allows RDF-style assertions about entities to be contextualized, and thus different values can be assigned for the same entity under distinct contexts, which is a crucial feature for modeling ecological as well as many other types of scientific data [6], [7]. In addition, OBOE is currently implemented as an OWL-DL ontology that can be easily used with (or extended by) other ontologies for specifying domain-specific types of entities, characteristics, measurement standards, protocols, and relationships. For instance, the Semtools project has defined specific OBOE extensions in collaboration with the Santa Barbara Coastal Long-Term Ecological Research Project as well as through ongoing collaborations with other projects, and general extensions exist for OBOE that define a number of common entities, measurement units, and corresponding physical characteristics.

B. Semantic Annotations

A semantic annotation consists of two parts: (i) a "configuration" of the observation model containing the specific entities, characteristics, observations, measurements, and so on (drawn from one or more domain ontologies) that appropriately capture the semantics of the data set; and (ii) a mapping between the attributes in the data set to specific measurements defined in the model configuration. Fig. 2 shows a high-level example of an annotation defined for a simple Kelp sampling data set. Here, the data set consists of five attributes (bottom of Fig. 2). Each attribute is mapped to a specific measurement type (where only the characteristic of each measurement type is shown), and measurement types are organized into observations of specific Kelp entities (shown of type "Macrocystis"), temporal points (denoted by date-times), and spatial locations (given as site names). Each measurement associated with a Kelp observation is assumed to have occurred within the site and during the given time as specified by the context relationships.

Semantic annotations can be used to facilitate discovery and integration of heterogeneous data sets. For instance, combining semantic annotations with OBOE, it is possible to discover data sets based on searches expressed over types of observations and measurements of interest. As simple examples, users can pose queries such as "*find all data sets containing observations of Kelp*" and "*find all data sets containing Mass measurements of Kelp*". Both of these queries would return the example data set in Fig. 2 since the attribute `WET` is linked to a WetMass measurement for observations of Macrocystis (where WetMass is defined as a special kind, or *subclass* of Mass, and Macrocystis is defined as a subclass of Kelp). Using semantic annotations in this way can help to increase both query recall and precision over standard keyword-based approaches [14]. In particular, by defining terms as subclasses of other terms (e.g., Macrocystis as a subclass of Kelp), term expansion can be used

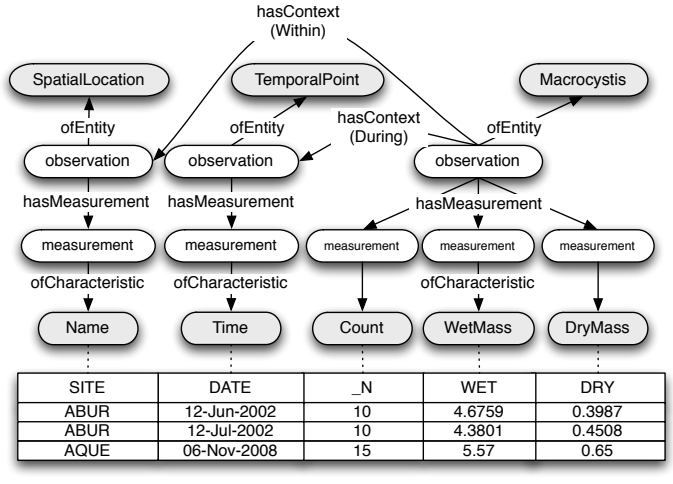

Fig. 2. Partial OBOE semantic annotation for Kelp sampling data. Shaded nodes represent ontological concepts; rectangular nodes are data table attributes mapped to OBOE measurement characteristics.

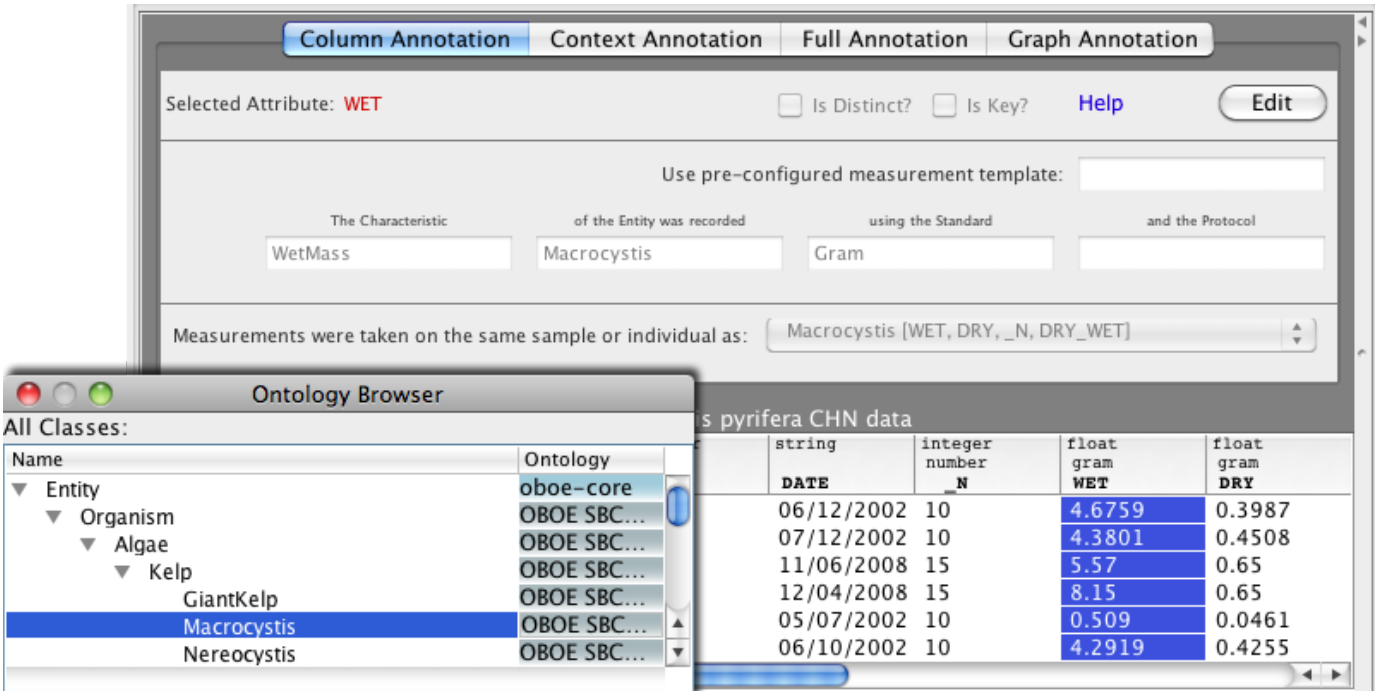

Fig. 3. Morpho metadata editor with Semantic plugin. The fill-in-the-blank interface uses natural language descriptions for intuitive editing. A searchable, hierarchical browser is used to select concepts from domain-specific ontologies.

to increase the number (recall) of data sets returned (where subclasses of query terms are also searched). The precision of the result can be improved since queries may specify the desired connections between terms (e.g., measurements of Mass for Kelp observations) as opposed to returning all data sets that simply mention the terms but without any explicit connections (i.e., where Mass was measured, but not for Kelp samples). Annotations also help facilitate integration by allowing tools to align data set attributes based on their declared measurement and observation types.

In general, semantic annotations provide a formal description of attribute semantics, whereas in many commonly-used metadata formats for describing data sets, only informal text-based descriptions of data attributes are permitted. In the case of EML, there is some overlap between these two mechanisms—particularly with respect to measurement standards—however, semantic annotations extend this approach by providing a general mechanism to formally associate concepts drawn from domain ontologies to attributes. In the Semtools framework, we employ an XML serialization syntax for semantic annotations that is compatible with EML but that is stored separately from the EML documentation of a data set (allowing, e.g., annotations to be used independently of EML or with other metadata standards if needed). In addition, semantic annotations can be used to "materialize" a given data set into a set of triples conforming to the model configuration given in the annotation. In other words, a tabular data set such as the one shown in Fig. 2 can automatically be converted into a corresponding collection of observation and measurement instances. This in turn enables a simple form of structural integration, where instead of having a large number of different tabular data structures, all data is represented using the standard set of structures defined by the OBOE model (see Fig. 1). Thus, materializing a data set in this way provides a more uniform structural representation that can make a number of discovery and integration tasks easier. For instance, materialization can be used to increase query expressivity by allowing searches of the form "*find all data sets containing Mass measurements of Kelp with values less than or equal to 5 grams*", which in our example can be answered by generating (i.e., materializing) the measurements associated with the `WET` and `DRY` attributes in the data set of Fig. 2.

C. The Semantic Mediation API

The Semantic Mediation API includes basic ontology management features, annotation manipulation capabilities, and simple concept navigation and visualization components. The API is intended to be a centralized toolkit for use in multiple application contexts (on either client or server deployments). The Semantic Mediation API uses both the OWL API [15] for ontology management services including ontology parsing, serialization, and simple class and property exploration as well as the Pellet description-logic reasoner [16] for classification and exposing inferred axioms in source ontologies. The inference services exposed through the Semantic Mediation API are used in both discovery and integration features described below. In our current Morpho and Metacat extensions, semantic annotations are managed and stored automatically in an underlying, local relational database. While it is also possible to use in-memory approaches for storing and querying annotations, we

found the overhead to be prohibitive when large numbers of data sets are managed.

D. The Morpho Editor Plugin

The semantic-annotation editor plugin for Morpho provides a front-end to the Semantic Mediation API and allows data owners and curators to define annotations for existing EML data descriptions. The editor provides a simple "fill-in-the-blank" style form-based interface with a searchable hierarchical concept selection widget (see Fig. 3). The plugin seamlessly integrates with a standard Morpho installation and provides semantic query capabilities for locating data packages, marking up data sets within a package using semantic annotations, and saving annotations locally or to a shared repository where they can be discovered and explored by other users. The annotation editor in Morpho allows a user to view the data set being annotated as they fill in (by selecting an appropriate ontology term) the characteristic, measurement standard, protocol, and associated entity for each data set attribute. Users can also specify whether an observation spans multiple columns, and can provide context relationships between attributes (i.e., observations). The editor provides a number of additional features including the ability to view the entire annotation (similar to Fig. 2) and to specify additional mapping constraints for observations and measurements.

E. Metacat Query Extensions

The semantic plugin for Metacat augments Metacat's existing metadata storage and search by allowing annotations to be saved and queried alongside the metadata and data that they annotate. In addition to traditional keyword and spatial search criteria, the Metacat plugin allows semantic criteria to be included where they may either increase query recall using term-expansion (i.e., traversing the class subsumption hierarchy) or refine the result set by limiting matches to data sets that contain the specified observational model (e.g., combinations of OBOE-compatible entity, characteristic, measurement standard, or protocol concepts). The observational model can be leveraged further by materializing the annotation and data artifact (via the Data Manager Library [13]) into a fully instantiated OBOE model and inspecting (and querying over) the observational values themselves.

III. Discovery and Integration

In this section we describe the new data discovery and integration applications we have built using the components described above as part of the Semtools project.

A. Concept Query

The semantic query interface (see Fig. 4) is implemented as a Web application over Metacat that primarily supports locating data sets by how well their observational models match the given criteria. The interface provides *structured* as opposed to *unstructured*, i.e., keyword-based queries. In particular, query criteria given by users largely mirror the structure of a semantic annotation in that combinations of Entity, Characteristic, and Protocol are specified and optionally compounded when increased precision is sought.

As discussed above, by leveraging the relationships defined and inferred from the ontology we are able to increase recall beyond what is possible for simple keyword-based searches [14]. Broad queries return direct matches as well as subclass matches. The queries can be quickly refined when using the Web application by allowing rapid exploration of the data repository without having to define complete observational queries *de novo*. The interface allows users to specify individual classes of a measurement as well as pre-configured measurement types (representing standard data set attribute types) as defined in OBOE compatible ontologies to enable a single concept to proxy its constituent parts, namely the characteristics of particular entities that can be measured with a set of protocols and standards. This short-hand query generation can save users time in specifying their queries, and highlights a compelling reason for using OBOE extension ontologies. Measurement templates can also be leveraged when creating or editing semantic annotations in the Morpho interface.

Using compound semantic query criteria applies a finer-grained filter on the data sets that are returned. Results can be restricted to only those data sets that include measurements for a set of specific characteristics of a particular observational entity. Furthermore, a query can specify that those measurements come from precisely the *same instance of that entity*; a feature that fully exercises the comprehensive observational structure expressed in the annotation and enables higher query precision as described above.

B. Data-Level Query

For even more precise recall, the OBOE model can be (partially) materialized (see above) during the query stage after which a data range filter can be applied. Different techniques are available for merging the annotation with the data that it describes, but for performance reasons a hybrid approach has been adopted in which preliminary search results from a structured query are collated and only that subset is materialized. Because our corpus is described using EML in conjunction with the annotation syntax, the Data Manager Library [13] is used to load the described raw data (into a relational database) while the annotation informs the correct use of the Data Manager query and filtering features. For any measurements that match the concept query criteria, we verify that those measurements (e.g., attributes) contain data values within the range specified in the initial semantic data query and return the data packages that contain them (see Fig. 4).

C. Data Integration

The materialization routine used for semantic data queries laid the groundwork for enabling data integration. In addition to inspecting the data for values within a range and returning the data sets that contain a match, the data integration feature of the Semtools Web application goes further by constructing a synthetic data product (table) that represents the complete results of the query in terms of both the attributes and the

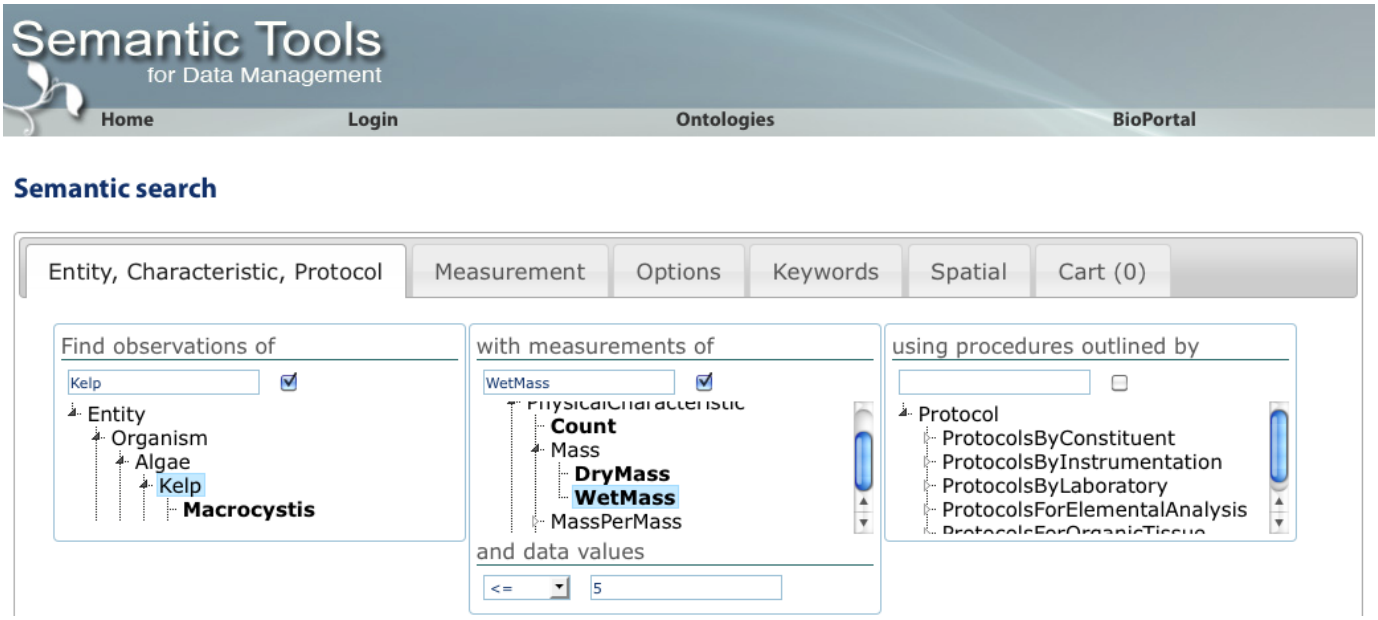

Fig. 4. Semantic data query web interface. Data packages containing observations of Kelp Wet Mass less than or equal to 5 [grams] are returned. Additional search options and compound query criteria can be specified within the other tabs. Matches can be saved in the data cart for later exploration.

values that are returned. Each original data set may have very different syntactic structures (e.g. column number, naming, order) but could share a subset of attributes that are semantically compatible as defined in accompanying annotations. These compatible, in-common attributes become the facets and metrics in a synthetic data set; their values filtered accordingly. Fig. 5 further illustrates the data integration support provided in the current implementation of the Web application. Consider two data packages, denoted by A and B in Fig. 5. Annotations (denoted C and D) are used to determine semantically equivalent data attributes contained in the data sets (denoted by E and F). The 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 parses each annotation and then computes an equivalence mapping among attributes based on their corresponding annotations. The Data Manager Library is then used to load the data sets and then query each data set to produce and merge the final, synthetic result data set.

While this approach provides a preliminary form of data-level integration, we are currently developing additional algorithms for determining compatibility of annotated measurements (e.g., to include unit information such as that gram and ounce are both mass units) and for converting measurement values using ontologically-defined unit conversions (e.g., 1000 milligrams in a gram), which will further support automated data integration through the Web application.

IV. Related Work

The need for more semantic mechanisms to describe observational data has led to many proposals for observational data models (e.g., [3], [5], [17]) and ontologies (e.g., [4], [6],

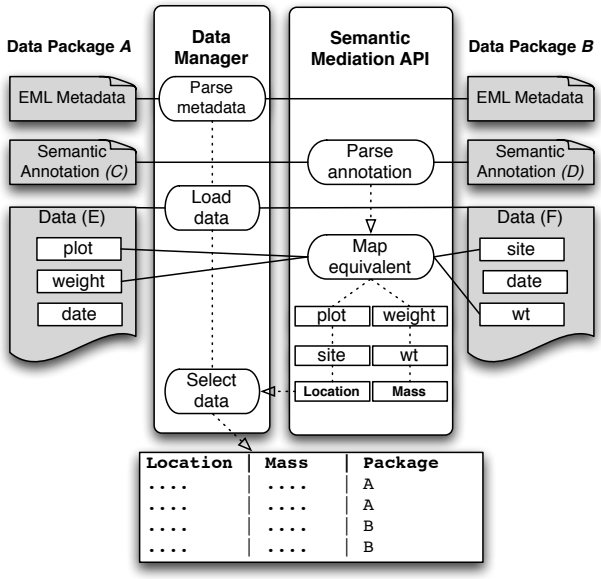

Fig. 5. Data integration query across multiple data packages (A, B). Annotations (C, D) are used to determine semantically equivalent data attributes contained in the data objects (E, F). Attributes `plot` and `site` are considered equivalent measurements of the characteristic Location; attributes `weight` 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.

[18]). The work presented here is complementary to these efforts by providing a concrete set of software components that have been integrated with popular metadata tools (namely, Metacat [11] and Morpho [12]) to provide a more uniform, semantic view of heterogeneous observational data. By ex-

tending Morpho and Metacat to support semantic annotations, these tools can provide additional help to researchers interested in performing synthetic studies by providing semantically-enhanced discovery and integration services, which are largely lacking in many existing environmental information management frameworks [19].

Our work on using semantic annotations for data integration is closely aligned to traditional information integration approaches (e.g., [20]), where a global mediated schema is used to (physically or logically) merge the structures of heterogeneous data sources using mapping constraints among the source and target schemas. As such, the observational model we employ in our framework can be viewed as a (general-purpose) mediation schema for observational data sets. This schema can be augmented with logic rules (as target constraints) where semantic annotations are used as mapping constraints. However, instead of users specifying logic constraints directly, we provide a high-level annotation language and user-interface components (through Morpho) that can simplify the specification of mappings and more naturally aligns with the observation model.

Annotations are playing a more prominent role in database systems, e.g., the MONDRIAN system [21] employs an annotation model and a set of query operators to manipulate both data and annotations. However, users must be familiar with the underlying data structures (schemas) to take advantage of these operators, which is generally not feasible for observational data in which data sets exhibit a high degree of structural and semantic heterogeneity. Our annotation approach used to extend EML is also similar in spirit to a number of other high-level mapping languages used for data exchange (e.g., [22], [23]). Our approach differs by being specifically tailored to the OBOE observational model, which in turn simplifies the annotation language, making it in general easier for users to specify annotations for observational data. Our approach also provides well-defined and unambiguous mappings from data sets to the observation and measurement model, which is critical for providing automated, high-quality data integration services over heterogeneous observational data.

V. Conclusion

The Semtools project has been successful in exploring and codifying technologies and techniques for applying semantic concepts to observational data. By providing mechanisms for linking data sets to ontological terms organized in a high-level observational model (e.g., OBOE), these new extensions to Metacat and Morpho help to overcome a number of limitations in existing metadata management systems that strive to provide effective data discovery and integration features. Our close involvement with the SONet Project (Scientific Observations Network) [24] encourages continued use-case refinement that will inform future semantic tool development and place an emphasis on intuitive interfaces and incremental adoption. This varied community of stakeholders is firmly invested in the use of cutting edge semantic solutions that will ultimately benefit multiple science disciplines by reducing obstacles to broad data sharing and innovative reuse.

Acknowledgement

This work supported in part through NSF grants 0743429 and 0753144.

References

[1] B. Worm, E. Barbier, N. Beaumont, J. Duffy, C. Folke, B. Halpern, J. Jackson, H. Lotze, F. Micheli, S. Palumbi, E. Sala, K. Selkoe, J. Stachowicz, and R. Watson, "Impacts of biodiversity loss on ocean ecosystem services," *Science*, vol. 314, no. 5800, pp. 787–790, 2006.

[2] S. Pennings, C. Clark, E. Cleland, S. Collins, L. Gough, K. Gross, D. Milchunas, and K. Suding, "Do individual plant species show predictable responses to nitrogen addition across multiple experiments?" *Oikos*, vol. 110, no. 3, pp. 547–555, 2005.

[3] OGC, "The OpenGIS Observations and Measurements Encoding Standard (O&M)," http://www.opengeospatial.org/standards/om.

[4] P. Fox, D. McGuinness, L. Cinquini, P. West, J. Garcia, J. Benedict, and D. Middleton, "Ontology-supported scientific data frameworks: The virtual solar-terrestrial observatory experience," *Computers & Geosciences*, vol. 35, no. 4, pp. 724–738, 2009.

[5] D. Tarboton, J. Horsburgh, and D. Maidment, "CUAHSI community observations data model (ODM), version 1.0 design specifications," http://water.usu.edu/cuahsi/odm/.

[6] C. Mungall, "Representing phenotypes in OWL," in *Proc. of the Workshop on OWL: Experiences and Directions (OWLED)*, 2007.

[7] S. Bowers, J. Madin, and M. Schildhauer, "A conceptual modeling framework for expressing observational data semantics," in *Proc. of the Intl. Conf. on Conceptual Modeling (ER)*, 2008, pp. 41–54.

[8] E. Fegraus, S. Andelman, M. Jones, and M. Schildhauer, "Maximizing the value of ecological data with structured metadata: An introduction to ecological metadata language (eml) and principles for metadata creation," *Bulletin of the Ecological Society of America*, vol. 86, no. 3, pp. 158–168, 2005.

[9] J. Madin, S. Bowers, M. Schildhauer, S. Krivov, D. Pennington, and F. Villa, "An ontology for describing and synthesizing ecological observation data," *Ecological Informatics*, vol. 2, no. 3, pp. 279–296, 2007.

[10] "OWL DL," http://www.w3.org/TR/owl2-overview/.

[11] D. Higgins, C. Berkley, and M. Jones, "Managing heterogeneous ecological data using morpho," in *Proc. of the Intl. Conf. on Scientific and Statistical Database Management (SSDBM)*, 2002, pp. 69–76.

[12] C. Berkley, M. Jones, J. Bojilova, and D. Higgins, "Metacat: A schema-independent XML database system," in *Proc. of the Intl. Conf. on Scientific and Statistical Database Management (SSDBM)*, 2001, pp. 171–179.

[13] B. Leinfelder, J. Tao, D. Costa, M. Jones, M. Servilla, M. O'Brien, and C. Burt, "A metadata-driven approach to loading and querying heterogeneous scientific data," *Ecological Informatics*, vol. 5, no. 1, pp. 3–8, 2010.

[14] C. Berkley, S. Bowers, M. Jones, J. Madin, and M. Schildhauer, "Improving data discovery for metadata repositories through semantic search," in *Proc. of the Intl. Conf. on Complex, Intelligent and Software Intensive Systems (CISIS)*, 2009, pp. 1152–1159.

[15] "The OWL API," http://owlapi.sourceforge.net/.

[16] Clark and Parisa, "Pellet: OWL 2 Reasoner for Java," http://clarkparsia.com/pellet/.

[17] Unidata, "network Common Data Form (netCDF)," http://www.unidata.ucar.edu/software/netcdf/.

[18] "Semantic Web for Earth and Environmental Terminology (SWEET)," http://sweet.jpl.nasa.gov/sweet/.

[19] M. Jones, M. Schildhauer, O. Reichman, and S. Bowers, "The new bioinformatics: Integrating ecological data from the gene to the biosphere," *Annual Review of Ecology Evolution and Systematics*, vol. 37, pp. 519–544, 2006.

[20] P. Kolaitis, "Schema mappings, data exchange, and metadata management," in *Symposium on Principles of Database Systems (PODS)*, 2005, pp. 61–75.

[21] F. Geerts, A. Kementsietsidis, and D. Milano, "Mondrian: Annotating and querying databases through colors and blocks," in *Proc. of the Intl. Conf. on Data Engineering (ICDE)*, 2006.

[22] R. Fagin, L. Haas, M. Hernandez, R. Miller, L. Popa, and Y. Velegrakis, "Clio: Schema mapping creation and data exchange," in *Conceptual Modeling: Foundations and Applications*, 2009, pp. 198–236.
[23] Y. An, J. Mylopoulos, and A. Borgida, "Building semantic mappings from databases to ontologies," in *Proc. of the AAAI*, 2006.
[24] "SONet: Scientific Observations Network," http://sonet.ecoinformatics.org/.