A Semantic Annotation Framework for Retrieving and Analyzing Observational Datasets

Shawn Bowers
Gonzaga University
bowers@gonzaga.edu

Huiping Cao *
New Mexico State University
hcao@cs.nmsu.edu

Mark Schildhauer
UC Santa Barbara
schild@nceas.ucsb.edu

Matt Jones
UC Santa Barbara
jones@nceas.ucsb.edu

Ben Leinfelder
UC Santa Barbara
leinfelder@nceas.ucsb.edu

Margaret O'Brien
UC Santa Barbara
mob@msi.ucsb.edu

ABSTRACT

In many scientific disciplines, including ecology, hydrology, and earth science, scientific analysis requires access to a broad range of observational data. However, because of the amount and heterogeneity (both in the structure and semantics) of observational data, approaches are needed that allow scientists to easily discover and analyze them. To address this issue, we describe a framework for accessing observational data. This framework combines a core observational model, domain-specific ontologies compatible with the core model, and a semantic annotation language. The annotation language provides a formal bridge between the core model and the underlying data to enable queries and analysis over annotations. The framework has been implemented to take advantage of ontology and web-based standards, and has also been integrated within a popular metadata tool for managing ecological datasets.

Categories and Subject Descriptors

H.4 [**Information Systems Applications**]: Miscellaneous

General Terms

Design

Keywords

Semantic annotation, Observational data

1. INTRODUCTION

Semantic annotations link one or more individual objects (e.g., all or a part of a text or an image) to semantic information that is often represented using formal ontologies [6]. This semantic information ideally helps users and systems interpret the object, which in turn can be used to improve the precision and recall of retrieval queries. However, effectively leveraging semantic annotations requires: (a) well-defined ontologies (with concepts and semantic relationships suitable for describing objects); (b) methods to annotate objects semantically (which often requires the ability to annotate specific parts of objects); and (c) algorithms to make use of semantic annotations, e.g., for improving information retrieval results. Over the past several years a number of large-scale ontology development efforts have been initiated with the goal of making it easier to discover and integrate scientific data. For instance, the OBO Foundry [1] has adopted several ontologies relevant to biological and biomedical research. In addition, several (semi-)automatic semantic annotation algorithms [6] have been proposed.

While semantic annotation approaches have provided a number of benefits in the molecular biology and biomedical domain [3], the same benefits are not as easily realized in other scientific fields such as ecology [5]. This is due to the fact that ecology is drawing together many types of information to address broad-scale questions about the natural world. Ecologists make extensive use of observational data, i.e., data that records observations and measurements (either by researchers or through, e.g., remote sensing) of real-world objects within different contexts such as space, time, etc. Although observational data is typically represented in a tabular form, i.e., rows and columns of data, a dataset contains a core set of canonical concepts [4, 2]: the *entities*, or objects, being observed; the *observations* of entities and their corresponding *measurements*; for each measurement, the value of a *characteristic* of the entity according to a measurement *standard* and *protocol* (or *procedure*); and the *context* assumed by each measurement and observation (e.g., the location where the entity was observed).

To fully achieve the benefits of data discovery and integration for ecological data it is crucial that semantic annotations are able to expose the above canonical concepts of datasets. For this purpose, we have developed the Extensible OBservations Ontology (OBOE) [4], which provides a common structural representation of the core concepts and their relationships. As an OWL ontology, OBOE can be easily extended to create domain-specific ontologies or used from within existing ontologies. OBOE is similar (and compatible) to a number of other efforts (e.g., the increasingly adopted O&M model developed through the OGC [2]) that have also acknowledged the importance of these canonical observation concepts for developing interoperable approaches for observational data. Based on the OBOE constructs, we provide a semantic annotation language that can be used to map underlying datasets to their corresponding observations and measurements. These semantic annotations together with domain-specific ontologies can be used for a variety of purposes. In what follows we briefly describe the OBOE ontology, our overall annotation framework, and our ongoing research in exploiting semantic annotations.

2. A CORE OBSERVATIONAL MODEL

Figure 1 shows the six major concepts represented by OBOE (here using standard UML syntax). Specifically, an Observation is made of some Entity (e.g., a biological organism or a geographic location), consists of zero or more Measurements of the Entity, and can be contextualized by zero or more other Observations.

*Most work was done when the author was in UC Santa Barbara.

ESAIR'10, October 30, 2010, Toronto, Ontario, Canada.
ACM 978-1-4503-0372-9/10/10.

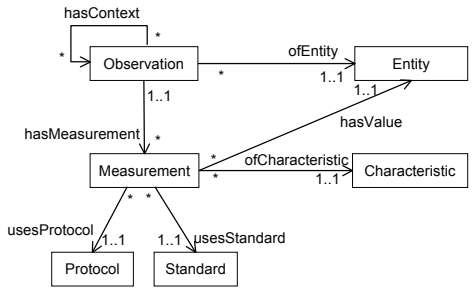

Figure 1: Basic concepts defined by OBOE.

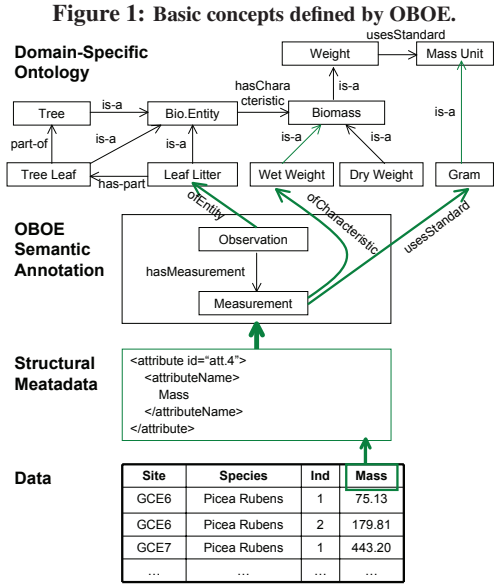

Figure 2: Semantic annotation framework based on OBOE.

Measurements assign values to Characteristics of Entities (e.g., the height or diameter of a tree), where a value can be another Entity (including primitive values such as integers and strings). Measurements also include Standards (e.g., units) and can specify Protocols (as well as other attributes, e.g., Precision, not shown in the figure).

In the example data of Figure 2, the first row shows that a scientist measures the mass (characteristic) of a "Picea Rubens" tree (entity), where the mass is recorded as 75.13 kg (value and standard). Another (simple) observation is made of the geographic site where the tree was observed. In this case, the name "GCE6" of the site was recorded. The observed "Picea Rubens" tree is contextualized by the observation of the site "GCE6". Here, the contextual relationship implies the tree was located within the site at the time of observation (this relationship can be explicitly established using OBOE, although not shown in the figure).

3. THE ANNOTATION FRAMEWORK

Figure 2 shows the different layers of our semantic annotation framework. We assume users first register their data (Data level) within a data repository. During annotation, one or more OBOE-compatible domain ontologies are selected. We assume ontologies are represented using the W3C OWL language, and we are currently developing tools within standard ontology editors to simplify the creation of OBOE-compatible ontologies. The user then selects the dataset and starts the annotation process. In general, each attribute of a dataset represents an OBOE measurement (with a characteristic, standard, etc.), and one or more dataset attributes represent an observation of an entity. We have developed a graphical editing tool for manually defining annotations, and are exploring approaches for semi-automatically generating annotations.

Figure 3 shows the high-level annotation syntax for the example

observation "o1"
  entity "Plot"
  measurement "m1" **key yes**
    characteristic "PlotName"
    standard "ManagedPlotCode"
observation "o2"
  entity "Tree"
  measurement "m2" **key yes**
    characteristic "TreeName"
    standard "TaxonomicName"
  measurement "m3" **key yes**
    characteristic "TreeLocalNo"
    standard "Nominal"
  measurement "m4"
    characteristic "Mass"
    standard "kg""
  context identifying yes "o1"
map "Site" to "m1"
map "Species" to "m2"
map "Ind" to "m3"
map "Mass" to "m4"

Figure 3: Semantic annotation

of Figure 2. OBOE observations and measurements are defined that are then mapped to specific dataset attributes. Also included are (optional) constraints for specifying details of the mapping.

Users can perform various operations over semantically annotated data without knowing the underlying data structures. As part of our current implementation, users can specify high-level discovery queries over the OBOE structure, and we have been able to show over a preliminary corpus that this approach can significantly improve query precision. We are also interested in leveraging annotations for helping users further refine and explore query results, including the following.

Summarizing data based on a given criteria. A natural next step after discoverying a data set is to perform simple summarizations, e.g., to determine the number of observations of a specific type, to find the ranges of certain measurements, and to display graphs of certain values. Summarizations can also be automated (e.g., based on context and other aspects of the annotation) and different data sets compared based on their summaries.

Mining and visualizing data to emphasize differences and similarities. A large number of search results are often returned for a given query, where many of the results are highly similar to each other. Our goal is to extend the summarization capabilities to further "classify" data into similar groups while highlighting differences within and among groups. In this way, we can alleviate the "too-many-answers" problem in answering simple queries, and help users to more easily narrow their searches (to find relevant data) without having to load and analyze each dataset separately.

4. CONCLUSIONS

We briefly described our work on using semantic annotation approaches to support a broad and important class of scientific data. Our approach leverages a core observational ontology and a declarative annotation syntax. Our annotation framework has been implemented within a common set of ecological data management tools. We described ongoing work to further leverage the framework for novel summarization and analytical tasks.

5. REFERENCES

[1] OBO foundry. http://www.obofoundry.org/.
[2] OpenGIS observations and measurements encoding standard (O&M) http://www.opengeospatial.org/standards/om.
[3] M. Ashburner, *et al*. Gene ontology: tool for the unification of biology. *Nat. Genet.*, 25:25–29, 2000.
[4] S. Bowers, J. S. Madin, and M. P. Schildhauer. A conceptual modeling framework for expressing observational data semantics. In *ER*, 2008.
[5] J. Madin, S. Bowers, M. Schildhauer, and M. Jones. Advancing ecological research with ontologies. *Trends in Ecology and Evolution*, 23(3):159–168, 2008.
[6] L. Reeve and H. Han. Survey of semantic annotation platforms. In *SAC'05: Proc. of the ACM symp. on Applied computing*, pages 1634–1638. ACM, 2005.