Protégé Extensions for Scientist-Oriented Modeling of Observation and Measurement Semantics (Application Track)

Wesley Saunders[1], Shawn Bowers[1], Margaret O'Brien[2]

[1]Department of Computer Science, Gonzaga University
[2]Marine Science Institute, UC Santa Barbara
wsaunders@zagmail.gonzaga.edu, bowers@gonzaga.edu, mob@msi.ucsb.edu

Abstract. We present Protégé-OWL extensions designed to help scientists define domain-specific ontologies for describing observational data. The extensions provide high-level forms that users can fill out from within Protégé to specify classes representing types of entities, characteristics (e.g., height, weight, area), standards (e.g., units), protocols, and context relationships used to describe scientific measurements. As a user fills out a form, the underlying OWL-DL axioms of the corresponding class are automatically asserted, thus allowing users to specify relatively complex OWL-DL constraints through the forms without requiring an understanding of the technical details of OWL. Encoded in the constraints generated by the extension are a set of "best practices" for enabling improved data discovery and integration of observational data.

1 Introduction

Earth and environmental scientists often depend on data collected from multiple research efforts to address broad and complex scientific questions. These efforts rely on effective approaches for discovering, interpreting, and integrating diverse and structurally heterogeneous data sets that cover a wide range of semantic concepts (including different geographic, temporal, and biological scales). Similar to molecular biology and biomedicine [1], employing domain-specific terms for describing earth and environmental data has the potential to significantly improve discovery and integration, however, only a relatively small number of ontologies have been created within these domains. We see two main barriers to ontology development within these communities: (1) the breadth of (specialized) concepts and phenomena studied within earth and environmental science requires a large and diverse number of ontological terms and ontologies, and (2) the high-level of expertise needed to efficiently develop ontologies using current ontology languages and tools.

The aim of this work is to help address these challenges by adding structured, easy-to-use forms to Protégé-OWL that scientists can use to quickly create meaningful domain-specific ontologies. Our approach leverages a generic, core observation and measurement ontology [2] that is designed for describing scientific data sets based on metadata annotations (mappings from data attributes to specialized measurement classes) [3]. These annotations provide a uniform view over otherwise heterogeneous data sets that can be used to enhance data discovery and integration applications (e.g., for improved precision and recall [3] or analysis over an integrated data repository [4]).

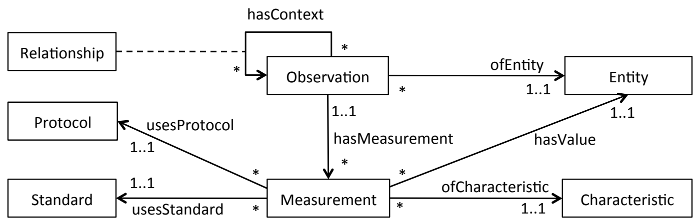

Fig. 1. The main classes and properties of the observation and measurement ontology (OBOE). Although shown in UML, the model is defined within OWL-DL.

This paper focuses on our extensions to Protégé-OWL that allow users to create sophisticated term definitions using simple "fill-in-the-blank" forms, while automatically generating the underlying DL axioms corresponding to the user's input. This approach shields users from having to understand the underlying DL formalisms, while still allowing axioms to be viewed using the standard Protégé interface. The axioms generated also encode a number of OWL-DL best practices (e.g., as in [5]) to ensure term definitions are well-suited for data discovery. We briefly describe the core observation and measurement ontology, the approaches used within our Protégé extensions, and conclude with a brief summary of our contributions and future work.

2 Observation Modeling using Protégé-OWL

Fig. 1 shows the main modeling constructs of OBOE[1] (the Extensible Observation Ontology) [2] used within our extensions. 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 height of a tree), and can also include *standards* (e.g., units) and collection *protocols*. An observation can occur within the surrounding *context* of other observations, where context can be viewed as a form of dependency [2], and context often includes a named relationship (e.g., "partOf", "within") between observed entities. 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 crucial for modeling scientific data [2,6].

Fig. 2 shows an example of the extensions (i.e., plug-in) we have developed within Protégé-OWL for defining domain-specific ontologies using OBOE. The extensions were developed as a new Tab within Protégé (version 4.1). When the OBOE Tab is initially selected from within the Tabs menu of Protégé the plug-in automatically imports the standard OBOE ontologies (after prompting the user). The OBOE Tab provides

[1] `http://ecoinformatics.org/oboe/oboe.1.0/oboe-core.owl`

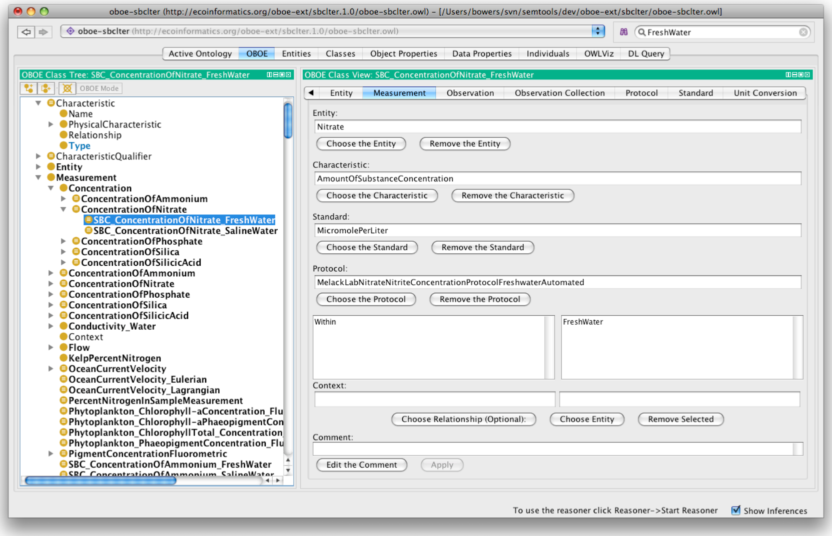

Fig. 2. The Measurement form showing the various fields of a "Fresh Water Nitrogen Concentration" measurement type defined within the an OBOE extension ontology.

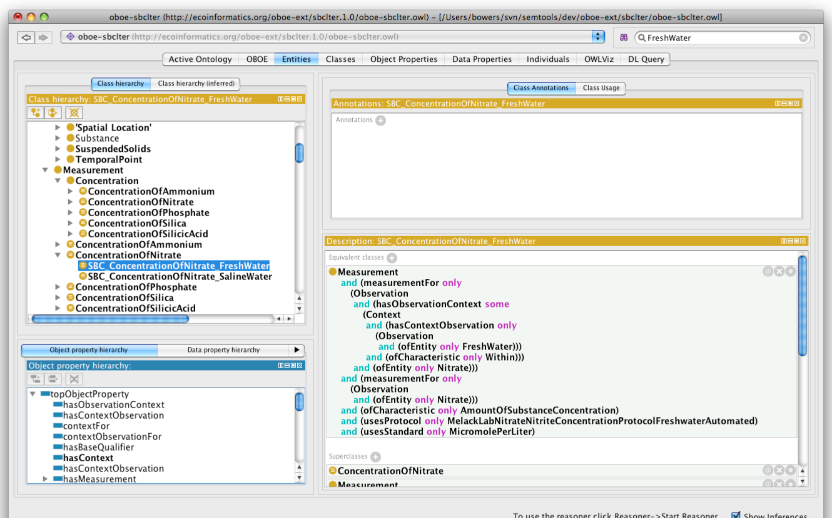

Fig. 3. The OWL-DL axioms created by the form in Fig. 2.

the standard class hierarchy view from Protégé (left-hand side of Fig. 2) as well as a new editing panel (right-hand side) with separate subpanels (subtabs) for each of the constructs shown in Fig. 1. Each subpanel contains a different form for specifying the associated OBOE class (the Measurement class form is shown in Fig. 2). Users create new classes via the standard class hierarchy panel. After a class is created or selected, the plug-in displays the appropriate form on the right-side of the window. Each form consists of a comment section as well as fields that can (optionally) be filled in. Most of these fields are filled with classes that are selected using a tree-based class selection widget, which constrains the choice of classes based on the type of class to be selected and the other values of fields as appropriate (e.g., depending on the characteristic chosen, only certain unit types can be selected). The Measurement form (shown in Fig. 2) contains the largest number of fields of all the forms, and includes fields for an observed entity, characteristic, standard, protocol, and zero or more context observations. Each context observation consists of an optional relationship type and an entity class (e.g., Fig. 2 shows the FreshWater entity and the Within relationship).

Fig. 3 shows the standard Protégé view for the class of Fig. 2. As shown, defining this class using the Measurement form results in a non-trivial DL axiom. In this case, we assert Measurement types (such as the one in Fig. 2) using an equivalent class axiom. A measurement type can be viewed as a combination of a number of other classes, and users can annotate data set attributes either directly via a measurement type or by specifying the individual components (i.e., the entity, characteristic, standard, and so on). By using equivalence classes, attributes can be classified into measurement types automatically using a reasoner (such as Pellet), which also allows for data discovery searches that are based either on measurement types or the individual components of a measurement. We note that most other classes created using the OBOE plug-in are defined using subclass axioms. As shown in Fig. 3 we also control the use of universal and existential property restrictions largely following the conventions defined in [5].

3 Summary and Future Work

We described new extensions to Protégé-OWL to help simplify the creation of scientific observation and measurement ontologies. The extensions are currently being applied to develop domain-specific controlled vocabularies of common measurement types used within the Santa Barbara Coastal Long-Term Ecological Research Project as well as those used within trait-based ecological and evolutionary research data (via the TraitNet project). These ontologies consist of thousands of terms created using the form-based approach described here. As future work, we are developing new features within the plug-in for extending context support and for modeling additional types of entity properties. We are also interested in extending our work here to support collaborative, web-based editing of observation and measurement ontologies.

References

1. B. Smith, *et al.*: The OBO Foundry: coordinated evolution of ontologies to support biomedical data integration. Nature Biotechnology **25** (2007) 1251–1255
2. Bowers, S., Madin, J., Schildhauer, M.: A conceptual modeling framework for expressing observational data semantics. In: ER. (2008)
3. Berkley, C., Bowers, S., Jones, M.B., Madin, J.S., Schildhauer, M.: Improving data discovery for metadata repositories through semantic search. In: CISIS. (2009) 1152–1159
4. Bowers, S., Kudo, J., Cao, H., Schildhauer, M.: ObsDB: A system for uniformly storing and querying heterogeneous observational data. In: e-Science. (2010)
5. A. Rector, *et al.*: OWL Pizzas: Practical experience of teaching OWL-DL: Common errors & common patterns. In: EKAW. (2004) 63–81
6. Mungall, C.: Representing phenotypes in owl. In: OWLED. (2007)