Towards a Generic Framework for Semantic Registration of Scientific Data*

Shawn Bowers and **Bertram Ludäscher**
San Diego Supercomputer Center
La Jolla, CA, 92093, USA
{bowers,ludaesch}@sdsc.edu

Introduction

Ontologies provide standard sets of terms and formal definitions for concepts in a domain, and are increasingly used to uniformly access information. For example, using concepts from an ontology to annotate disparate information (such as Web-page content) allows ontology-aware applications to retrieve and navigate otherwise heterogeneous sources.

In this paper, we consider the specific problem of registering scientific data (as opposed to arbitrary Web content) with ontologies. We propose a generic framework to support *semantic registration* of scientific datasets, which we intend to deploy in the SEEK[1] project—a multidisciplinary effort to help scientists discover, access, integrate, and analyze distributed ecological information. Our goal is to develop a framework that can support data providers by allowing them to easily enrich their datasets using ontological concepts, while at the same time providing end users semantic-discovery and data-mediation services.

To illustrate the goal of semantic registration, Figure 1 gives an example ecological ontology that includes concepts for datasets, measurements, and locations. The ontology is expressed using standard description-logic syntax (Sattler 2003), but could just as easily be expressed using OWL (McGuinness & van Harmelen 2003). Figure 2 shows part of a dataset,[2] which includes concepts from Figure 1. In particular, the dataset has abundance data for plant and animal species located in giant-kelp forests along specific regions of the Santa Barbara Long-Term Ecological Research (LTER) site. The dataset gives the date and location of samples, species codes, and the number of species observed. Through semantic registration it should be possible to connect the dataset of Figure 2 to the appropriate portions of the ontology of Figure 1.

The rest of this paper describes our proposed framework for semantic registration. The framework lets a data provider choose the appropriate concepts within an ontology that best describe the dataset. Based on the data provider's selections, a semantic registration tool can make suggestions for structural representations of the selected concepts. The data provider can then fit the structural representation (i.e., conceptual schema) to the dataset by specifying the connection between the information within the dataset and the structures of the schema. We describe the basic components of the framework in Section 2 and conclude in Section 3 by discussing some of the issues raised by our framework.

A Framework for Semantic Registration

The role of a *data provider* is to register a particular dataset with the appropriate ontology (or possibly multiple ontologies). Semantic registration occurs in two steps, as shown in Figure 3.

First, the data provider selects the concepts (and possibly the roles) in the ontology that are relevant to the dataset. We intend for the ontology to be presented graphically to the user, e.g., as a simple hierarchy of terms. As an example, the data provider might choose the concepts SpeciesAbundanceMeasurement and SBLTERSite as relevant to the dataset of Figure 2. Based on the data provider's selections, the *conceptual-schema generator* suggests a corresponding logical schema that contains the necessary structures to classify the dataset. We note that in our framework, the ontology serves as an abstract, intensional definition of the concepts and roles in a domain—whereas the conceptual schema provides a concrete, structural definition of the classes and relationships needed to realize the abstract definitions. For example, Figure 4 gives one possible schema for the selected concepts of the dataset of Figure 2. To generate this conceptual schema, the conceptual-schema generator must be able to reason over the ontology, e.g., to compute roles that are inferred through inheritence definitions. Note that the process of generating the conceptual schema is (potentially) incremental, i.e., the user may select relevant concepts, then decide that additional concepts are required or not needed, and so on. Once the data provider determines the appropriate conceptual schema, the mapping used to generate the schema is stored. The mapping is used to support query and navigation of the dataset using the ontology.

The second task of the data provider is to define a *semantic mapping*, which describes the correspondence between information in the dataset and objects in a conceptual-

*This work supported in part by the National Science Foundation (NSF) under Grant No. ITR 0225676 (SEEK) and ITR 0225673 (GEON).

[1]The Science Environment for Ecological Knowledge, http://seek.ecoinformatics.org

[2]Taken from the Santa Barbara Long-Term Ecological Research Site, http://sbc.lternet.edu/data/CRSData.html.

$$
\begin{array}{rcl}
\text{DataCollectionEvent} & \equiv & \exists \text{contains.Measurement} \\
\text{Measurement} & \equiv & \exists \text{measureOf.MeasurableItem} \\
\text{FieldCollectionMeasurement} & \equiv & \text{Measurement} \sqcap \exists \text{hasTime.DateTime} \sqcap \exists \text{hasLoc.Location} \\
\text{MeasurableItem} & \sqsubseteq & \exists \text{hasUnit.Unit} \sqcap \exists \text{hasValue.UnitValue} \\
\text{SpeciesCount} & \sqsubseteq & \text{MeasurableItem} \sqcap \exists \text{hasSpecies.Species} \sqcap \exists \text{hasUnit.RatioUnit} \\
\text{AbundanceMeasurement} & \sqsubseteq & \text{FieldCollectionMeasurement} \\
\text{SpeciesAbundanceMeasurement} & \sqsubseteq & \text{AbundanceMeasurement} \sqcap \exists \text{measureOf.SpeciesCount} \\
\text{FieldCollectionEvent} & \sqsubseteq & \text{DataCollectionEvent} \sqcap \exists \text{contains.FieldCollectionMeasurement} \\
\text{AbundanceCollectionEvent} & \sqsubseteq & \text{FieldCollectionEvent} \sqcap \exists \text{contains.SpeciesAbundanceMeasurement} \\
\text{Location} & \equiv & \exists \text{position.Coordinate} \\
\text{LTERSite} & \sqsubseteq & \text{Location} \\
\text{SBLTERSite} & \sqsubseteq & \text{LTERSite} \\
\{\text{naples}\} & \sqsubseteq & \text{SBLTERSite}
\end{array}
$$

Figure 1: Example of an ecological ontology.

Date	*Site*	*Transect*	*SP Code*	*Count*
2000-09-08	CARP	1	CRGI	0
2000-09-08	CARP	4	LOCH	0
2000-09-08	CARP	7	MUCA	1
2000-09-22	NAPL	7	LOCH	1
2000-09-18	NAPL	1	PAPA	5
2000-09-28	BULL	1	CYOS	57

Figure 2: A partial dataset for invertebrate and algae counts in 20x1 meter-square quadrats.

schema instance. The conceptual-schema instance is called the *object-base* in Figure 3. In particular, a semantic mapping should define the *logical objects* in the dataset (like specific abundance measurements or specific species) along with the *logical relationships* among objects with respect to the conceptual schema (e.g., that a particular abundance measurement was for a specific species).

In our framework, we envision a registration mapping as a set of logic rules expressed using Datalog (Abiteboul, Hull, & Vianu 1995). (Note that these rules would be generated from a high-level language, e.g.) Thus, the object-base may not be explicitly instantiated, and instead serves as a virtual extent that can be populated on demand. The use of logic rules offers a flexible, and declarative language for expressing registration mappings. For example, the following rules define a mapping between the dataset of Figure 2 and the conceptual schema of Figure 4. We use RDF(S) (Lassila & Swick 1999) as the language for the object-base and conceptual schema, respectively. An RDF triple is represented using the formula *rdf(S,P,V)*, where S, P, and V denote the subject, predicate, and value of the triple. The formula *idGen*(L,I) represents a Skolem function for mapping a list L of atoms to an identifier I. The formula *anonGen*(L,I) is similar to *idGen*, but generates an anonymous resource I based on the atoms of L. The dataset has the identifier "sb.979.7" and is accessed (via the *format wrapper*) using the formula $t(Date, Site, Transect, SPCode, Count)$.

```
rdf(sb.878.7,rdf:type,'AbundanceCollection').
rdf(I,rdf:type,'SpeciesAbundance') :-
   t(D,S,T,C,_), idGen([D,S,T,C],I).
rdf(sbclter.878.7,contains,I) :-
   t(D,S,T,C,_), idGen([D,S,T,C],I).
```

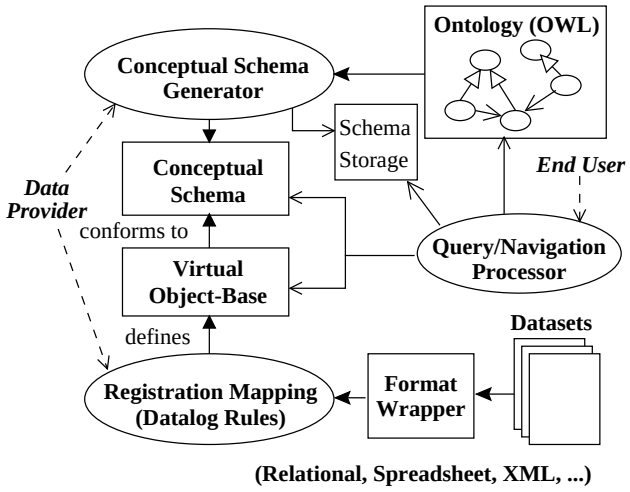

Figure 3: The proposed semantic registration framework.

```
rdf(A,rdf:type,'SpeciesCount') :-
   t(D,S,T,C,_), anonGen([D,S,T,C],A).
rdf(I,measureOf,A) :-
   t(D,S,T,C,_), idGen([D,S,T,C],I),
   anonGen([D,S,T,C],A).
rdf(I,locOf,'Naples') :-
   t(D,'NAPL',T,C,_),
   anonGen([D,'NAPL',T,C],A).
rdf(I,timeOf,D) :-
   t(D,S,T,C,_), anonGen([D,S,T,C],A).
rdf(A,hasValue,N) :-
   t(D,S,T,C,N), anonGen([D,S,T,C],A).
rdf(A,hasSpecies,'crassedoma-giganteum') :-
   t(D,S,T,'CRGI',_),
   anonGen([D,S,T,'CRGI'],A).
```

Once a dataset is registered, an *end user* can query and navigate it using the concepts and roles in an ontology. For example, a scientist may be interested in finding all LTER research sites that have abundance data on crassedoma giganteum (i.e., rock scallops). The query can be expressed using the ontological concepts LTERSite, AbundanceMeasurement, and Species. To answer the query, the appropriate mapping between the ontology and the conceptual-

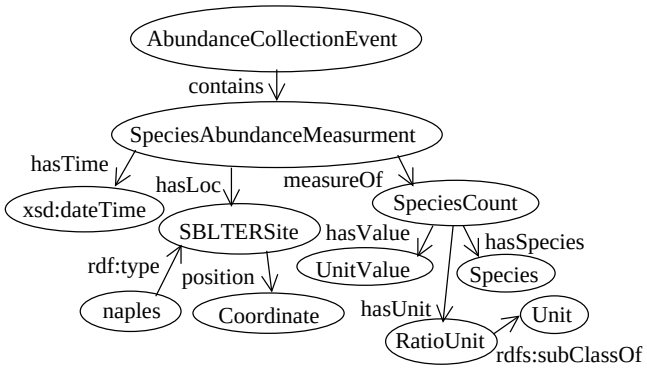

Figure 4: A conceptual schema generated from a subset of the ecological ontology.

schema must be retrieved, which is then used along with the registration mapping to determine the corresponding LTER sites. Note that, similar to constructing a conceptual schema, a reasoner is needed to match ontological concepts of the query with the associated conceptual schemas of the registered datasets.

Open Issues

The framework described in the previous section presents a number of open issues and technical questions, some of which we summarize below.

- *What interaction is needed between the conceptual schema generator and the data provider?* In particular, the desired outcome of semantic mapping is to connect a relational table to the conceptual schema. At issue is whether there is a need to restructure the generated schema, e.g., to better suit the dataset being registered or to make semantic mappings easier to specify. Is there a single conceptual schema for the ontology(ies), or are slightly different schemas required to make semantic mappings (from datasets) easier to specify?
- *What inference procedures are required to generate conceptual schemas from ontologies?* Depending on the interaction needed and the expressive power of the ontology language, is it possible to use existing description-logic reasoners, e.g., FaCT (Horrocks 1999), to generate schemas, or are special-purpose reasoning systems needed? Which approach is better? Similarly, what reasoning power is needed to match queries with corresponding conceptual schemas?
- *Is RDF(S) an appropriate object-model for semantic registration?* We want to use, when possible, standard Semantic Web languages in SEEK. However, is RDF(S) and OWL sufficient for semantic registration? Is it reasonable to expect data providers to use RDF(S) to generate registration mappings? If not, what high-level languages and interfaces can be used on top of RDF(S) and OWL that will enable semantic registration? Similarly, can languages such as Prolog, Datalog, or F-Logic (Kifer & Lausen 1989) be practically used to support semantic views over datasets?
- *What should happen when constraints in the ontology are violated by the dataset?* Datasets provide a subset of a domain, e.g., a dataset may have locations but not coordinates. However, coordinates are essential to locations in the ontology. One approach is to "fill-out" registration mappings, e.g., by using geographic services that take transects within LTER sites and return coordinates.
- *Is it possible to use high-level languages to specify registration mappings? Is it possible to automate portions of the registration mapping?* Can metadata or dataset constraints be used to generate mappings? Alternatively, are there appropriate high-level languages for data providers to express mappings? And, can mapping rules be generated from the specification?
- *Is the proposed framework suitable for semantic mediation?* Semantic registration is an essential component of a semantic mediation system. One issue is whether our framework is suitable for enabling semantic mediation in the presence of incremental dataset discovery (Gupta, Ludäscher, & Martone 2002). In addition, what is needed from our framework to retrieve data from multiple datasets through queries expressed against the ontologies? And how should this retrieved data be represented (e.g., using the intermediate conceptual schema)?

References

[Abiteboul, Hull, & Vianu 1995] Abiteboul, S.; Hull, R.; and Vianu, V. 1995. *Foundations of Databases*. Addison-Wesley Publishing.

[Gupta, Ludäscher, & Martone 2002] Gupta, A.; Ludäscher, B.; and Martone, M. E. 2002. Registering scientific information sources for semantic mediation. In *Proceedings of the 21st International Conference on Conceptual Modeling (ER)*, number 2503 in Lecture Notes in Computer Science, 182–198. Springer-Verlag.

[Horrocks 1999] Horrocks, I. 1999. FaCT and iFaCT. In *The International Workshop on Description Logics*, CEUR Workshop Proceedings. Linköping.

[Kifer & Lausen 1989] Kifer, M., and Lausen, G. 1989. F-Logic: A higher-order language for reasoning about objects, inheritance, and scheme. In *Proceedings of the 1989 ACM SIGMOD International Conference on Management of Data*, 134–146. ACM Press.

[Lassila & Swick 1999] Lassila, O., and Swick, R. R., eds. 1999. *Resource Description Framework (RDF) Model and Syntax Specification*. W3C Recommendation. World Wide Web Consortium (W3C). http://www.w3.org/TR/1999/REC-rdf-syntax-19990222/.

[McGuinness & van Harmelen 2003] McGuinness, D. L., and van Harmelen, F., eds. 2003. *OWL Web Ontology Language Overview*. W3C Candidate Recommendation. World Wide Web Consortium (W3C). http://www.w3.org/TR/2003/CR-owl-features-20030818/.

[Sattler 2003] Sattler, U. 2003. Description logics for ontologies. In *Proceedings of the International Conference on Conceptual Structures (ICCS)*, volume 2746 of *Lecture Notes in Artificial Intelligence*. Springer-Verlag.