Data Procurement for Enabling Scientific Workflows: On Exploring Inter-ant Parasitism*

Shawn Bowers[1], David Thau[2], Rich Williams[3], and Bertram Ludäscher[1]

[1] San Diego Supercomputer Center, UCSD, La Jolla, CA, USA
[2] University of Kansas, Lawrence, KS, USA
[3] National Center for Ecological Analysis and Synthesis,
UCSB, Santa Barbara, CA, USA

1 Introduction

Similar to content on the web, scientific data is highly heterogeneous and can benefit from rich semantic descriptions. We are particularly interested in developing an infrastructure for expressing explicit semantic descriptions of ecological data (and life-sciences data in general), and exploiting these descriptions to provide support for automated data integration and transformation within scientific workflows [2]. Using semantic descriptions, our goal is to provide scientists with: (1) tools to easily search for and retrieve datasets relevant to their study (i.e., data *procurement*), (2) the ability to select a subset of returned datasets as input to a scientific workflow, and (3) automated integration and restructuring of the selected datasets for seamless workflow execution.

As part of this effort, we are developing the *Semantic Mediation System* (SMS) within the SEEK project[1], which aims at combining knowledge representation and semantic-web technologies (e.g., OWL and RDF) with traditional data-integration techniques [3, 8, 9]. We observe that along with these traditional approaches, mediation of ecological data also requires external, special-purpose services for accessing information not easily or conveniently expressed using conceptual modeling languages, such as description logics. The following are two specific examples of ecologically relevant, external services that can be exploited for scientific-data integration and transformation.

Taxonomic Classification and Mapping. There is an extensive body of knowledge on species (both extinct and existing) represented in a variety of different taxonomic classifications, and new species are still being discovered [7]. The same species can be denoted in many ways across different classifications, and resolving names of species requires mappings across multiple classification hierarchies [6]. Within SMS we want to leverage operations that exploit these existing mappings, e.g., to obtain synonyms of species names, without explicitly representing the mappings or simulating the associated operations within the mediator.

Semantics-Based Data Conversion. We are interested in applying operations during mediation that can transform and integrate data based on their implied meaning. How-

* This work supported in part by NSF grant ITR 0225674 (SEEK).

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

C. Bussler et al. (Eds.): SWDB 2004, LNCS 3372, pp. 57–63, 2005.

ever, for scientific data, the nature of these conversions are often difficult to express explicitly within a conceptual model. A large number of ecological datasets represent real-world observations (like measuring the abundance of a particular species), and therefore often have slightly different spatial and temporal contexts, use different measurement protocols, and measure similar information in disparate ways (e.g., area and count in one dataset, and density, which is a function of area and count, in a second dataset). As with taxonomic classification, we want the mediator to exploit existing conversion operations when possible.

This short paper describes an initial logic-based SMS prototype that leverages ontologies, semantic descriptions, and simple external services (primarily taxonomic) to help researchers find relevant datasets for ecological modeling. In Section 2 we describe our motivating scenario. In Section 3 we discuss details of the prototype through examples. And in Section 4 we conclude with future work.

2 Motivation: Ant Parasitism and Niche Modeling

A diverse and much studied group of organisms in ecology is the family *Formicidae*, commonly known as ants. Ants account for a significant portion of the animal biomass on earth and churn much of the earth's soil. Ants are also social animals that provide insights into the evolution of social behaviors. One such complex social behavior is parasitism between ant species [4].

The environment in which parasitism is likely to occur provides important data on how parasitism arises. For example, one theory states that inter-ant parasitism is more likely to arise in colder climates than in warmer ones. Thus, an ecological researcher may be interested in the question: *In California, what environmental properties play a role in determining the ranges of ants involved in inter-ant parasitism?*

Answering this question requires access to a wide array of data: (1) the types of parasitic relationships that exist between ants, (2) the names of species of ants taking part in these parasitic relationships, (3) georeferenced observations of these species of ants, and (4) the climate and other environmental data within the desired locations.

Today, these datasets are typically sought out by the researcher, retrieved, and integrated manually. The researcher analyzes the data by running it through an appropriate ecological model, the result of which is used to help test a hypothesis. In our example, an ecological niche model [10] can be used, which takes data about the presence of a species and the environmental conditions of the area in question, and produces a set of rules that define a "niche" (i.e., the conditions necessary for the species to exist) relative to the given environmental conditions and presence data. The rest of this paper describes a first step towards helping a researcher to collect the datasets needed to test inter-ant parasitism, and similar high-level questions.

3 The Prototype

Our dataset-discovery architecture is shown in Figure 1. A set of repositories store ontological information, datasets, and semantic descriptions (of the datasets). A semantic

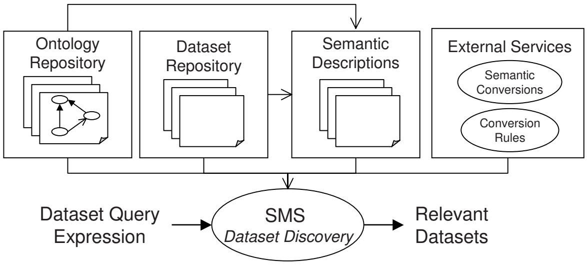

Fig. 1. The initial SMS architecture for ecological data mediation

d_1

genus	*species*	*count*	*lat*	*lon*
Manica	parasitica	2	37.85	-119.57
Manica	bradelyi	1	38.32	-119.67

d_2

genus	*species*	*cnt*	*lt*	*ln*
Camponotus	fornasinii	1	-29.65	26.18

d_3

man-para-cnt	*aph-cald-cnt*	*lt*	*ln*
3	6	37.56	-120.03

d_4

genus1	*species1*	*genus2*	*species2*
Manica	parasitica	Aphaenogaster	calderoni

Fig. 2. Four heterogeneous datasets d_1 through d_4

description logically annotates a dataset using concepts and roles in the ontology repository. Semantic descriptions are expressed as sound *local-as-view* mappings [3, 8], which can succinctly represent mappings from information within a dataset to corresponding ontological information. We also consider external services in the architecture, which currently consist of synonym and unit-conversion operations. The SMS engine accepts a user query and returns the set of relevant datasets that satisfy the given query.

Figure 2 shows example portions of four datasets that can be used to help answer ant and inter-ant parasitism queries. Dataset d_1 in Figure 2 contains georeferenced ant data from AntWeb[2] and consists of approximately 1,700 observations, each of which consist of a genus and species scientific name, an abundance count, and the location of the observation. Dataset d_2 in Figure 2 contains similar georeferenced ant data from the Iziko South African Museum (ISAM),[3] consisting of about 12,000 observations. Dataset d_3 in Figure 2 is a typical representation used for georeferenced co-occurrence data, where species are encoded within the schema of the table. This dataset contains only five tuples. Dataset d_4 in Figure 2 describes specific ants that participate in inquilinism inter-ant parasitism. The first two columns denote the parasite and the last two columns denote the host. Over two-hundred pairs of ants are described using four distinct datasets, each representing a particular parasitic relationship (all data were derived from Table 12-1 of [4]). Finally, Figure 3 shows a simplified fragment of the measurement and parasitism ontologies currently being developed within SEEK.

The following conjunctive queries define semantic descriptions of datasets d_1, d_3, and d_4 (the semantic description of d_2 is identical to d_1).

[2] See www.antweb.org

[3] Provided by Hamish Robertson, Iziko Museums of Cape Town

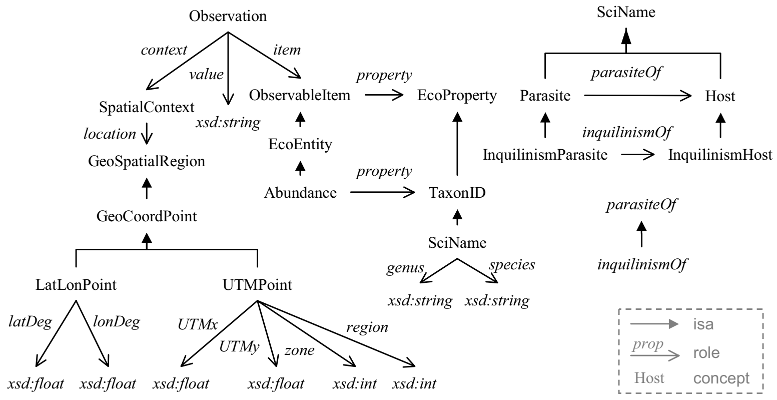

Fig. 3. Simplified ontologies for measurement observations and inter-ant parasitism

d_1(Ge,Sp,Co,Lt,Ln) :-
Observation(O), value(O,Co), context(O,S), location(S,P), LatLonPoint(P), latDeg(P,Lt), lonDeg(P,Ln), item(O,A), Abundance(A), property(A,N), SciName(N), genus(N,Ge), species(N,Se).
d_3(Mp, Cf, Lt, Ln) :-
Observation(O_1), value(O_1,Mp), context(O_1,S), location(S,P), LatLonPoint(P), latDeg(P,Lt), lonDeg(P,Ln), item(O_1,A_1), Abundance(A_1), property(A_1,N_1), SciName(N_1), genus(N_1,‘Manica’), species(N_1,‘parasitica’), Observation(O_2), value(O_2,Cf), context(O_2,S), item(O_2,A_2), Abundance(A_2), property(A_2,N_2), SciName(N_2), genus(N_2,‘Aphaenogaster’), species(N_2,‘calderoni’).
d_4(G_1,S_1,G_2,S_2) :-
InquilinismParasite(P), SciName(P), genus(P,G_1), species(P,S_1), InquilinismHost(H), genus(H,G_2), species(H,S_2), inquilinismOf(P,H).

The following example is a dataset-discovery query defined in terms of the ontology that asks for all datasets containing georeferenced abundance measurements of Manica bradleyi ants observed within California (as defined by the given bounding box). Dataset-discovery queries allow predicates to be annotated with dataset variables, given as D below. Each semantic description is also implicitly annotated with its dataset identifier, e.g., every predicate in the body of the first description above would be annotated with d_1. A dataset handle is returned by the query below if each formula annotated with D is satisfied by the dataset, assuming the given inequality (i.e., the latitude-longitude) conditions also hold.

q_1(D) :- Observation(O)^D, context(O,S)^D, location(S,P)^D, LatLonPoint(P)^D, latDeg(P,Lt)^D, lonDeg(P,Ln)^D, item(O,A)^D, Abundance(A)^D, property(A,N)^D, SciName(N)^D, genus(N,‘Manica’)^D, species(N,‘bradleyi’)^D, Lt $\geq$ 33, Lt $\leq$ 42, Ln $\geq$ -124.3, Ln $\leq$ -115.

Using a standard data-integration query-answering algorithm [8], the query above is answered by (1) finding *relevant* information sources, i.e., sources whose view mappings overlap with the given query, and (2) using the relevant sources, rewriting the user query, producing a sound query expressed only against the underlying data sources, possibly containing additional conditions. We extend this approach by also considering dataset annotations on query formulas. In our example, d$_1$ and d$_2$ are the only relevant datasets for the above query, giving the following query rewritings. Note that after executing the queries below, only d$_1$ is returned; the ISAM dataset does not contain the given species.

q$_1$(d$_1$) :- d$_1$(‘Manica’,‘bradleyi’,Ct,Lt,Ln), Lt $\geq$ 33, Lt $\leq$ 42, Ln $\geq$ -124.3, Ln $\leq$ -115.
q$_1$(d$_2$) :- d$_2$(‘Manica’,‘bradleyi’,Ct,Lt,Ln), Lt $\geq$ 33, Lt $\leq$ 42, Ln $\geq$ -124.3, Ln $\leq$ -115.

The following query is similar to q_1, but uses an external service (prefixed with ‘ext:’) for computing synonymy of species names.

q$_2$(D) :- Observation(O)^D, context(O,S)^D, location(S,P)^D, LatLonPoint(P)^D,
latDeg(P,Lt)^D, lonDeg(P,Ln)^D, item(O,A)^D, Abundance(A)^D, property(A,N)^D,
SciName(N)^D, genus(N,Ge)^D, species(N,Sp)^D, Lt $\geq$ 33, Lt $\leq$ 42, Ln $\geq$ -124.3,
Ln $\leq$ -115, ext:synonym(‘Manica’,‘bradleyi’,Ge,Sp).

The synonymy operation, encapsulated as a logical formula above, draws from descriptions in the Hymenoptera Name Server [5], and supports over 2,500 taxa of ants and their synonymy mappings. In the operation, a given genus-species pair is always a synonym of itself. In the prototype, we equate synonyms between taxa as equivalence relations. This assumption is often an oversimplification [1] and in future work we intend to explore different synonymy relations between taxa.

The following rewritings are obtained from the above query. After execution, the rewritten q_2 query will return dataset d$_1$ as well as dataset d$_3$; the latter because Aphaenogaster calderoni is a synonym of Manica bradleyi. Note that we could have discarded the third rewriting below since all arguments of the synonym operation are ground, and for the particular binding, the species’ are not valid synonyms.

q$_2$(d$_1$) :- d$_1$(Ge,Sp,Ct,Lt,Ln), Lt $\geq$ 33, Lt $\leq$ 42, Ln $\geq$ -124.3, Ln $\leq$ -115,
ext:synonym(‘Manica’,‘bradleyi’,Ge,Sp).
q$_2$(d$_2$) :- d$_2$(Ge,Sp,Ct,Lt,Ln), Lt $\geq$ 33, Lt $\leq$ 42, Ln $\geq$ -124.3, Ln $\leq$ -115,
ext:synonym(‘Manica’,‘bradleyi’,Ge,Sp).
q$_2$(d$_3$) :- d$_3$(Mp,Cf,Lt,Ln), Lt $\geq$ 33, Lt $\leq$ 42, Ln $\geq$ -124.3, Ln $\leq$ -115,
ext:synonym(‘Manica’,‘bradleyi’,‘Manica’,‘parasitica’).
q$_2$(d$_3$) :- d$_3$(Mp,Cf,Lt,Ln), Lt $\geq$ 33, Lt $\leq$ 42, Ln $\geq$ -124.3, Ln $\leq$ -115,
ext:synonym(‘Manica’,‘bradleyi’,‘Aphaenogaster,‘calderoni’).

Finally, the following query finds datasets containing georeferenced measurements of parasites of Manica bradleyi within California. Thus, the query finds the relevant ant presence data needed for our original parasitism question, for a single host species. The query uses the external synonym operation and projects the latitude, longitude, and genus and species names of the relevant observations so that the result (with additional pre-processing) can be fed into a scientific workflow, such as a niche model.

q$_3$(D,Lt,Ln,Ge,Sp) :- Observation(O)^D, context(O,S)^D, location(S,P)^D, LatLonPoint(P)^D,
latDeg(P,Lt)^D, lonDeg(P,Ln)^D, item(O,A)^D, Abundance(A)^D, property(A,N)^D,
SciName(N)^D, genus(N,Ge)^D, species(N,Sp)^D, Lt $\geq$ 32, Lt $\leq$ 42, Ln $\geq$ -124.3,
Ln $\leq$ -115, Host(Ho), genus(Ho,Ge$_1$), species(Ho,Sp$_1$),
ext:synonym(‘Manica’,‘bradleyi’,Ge$_1$,Sp$_1$), Parasite(Pa), genus(Pa,Ge$_2$),
species(Pa,Sp$_2$), parasiteOf(Pa,Ho), ext:synonym(Ge$_2$,Sp$_2$,Ge,Sp).

The rewritings of q_3 are shown below. The result includes the tuples (d$_1$,37.85,-119.57,‘Manica’,‘parasitica’) and (d$_3$,37.56,-120.03,‘Manica’,‘parasitica’), where only datasets d$_1$ and d$_3$ contain possible answers. In particular, Manica parasitica are inquilinism parasites of Manica bradleyi, which is derived from dataset d$_4$ by computing Manica bradleyi synonyms.

q$_3$(d$_1$,Lt,Ln,Ge,Sp) :- d$_1$(Ge,Sp,Ct,Lt,Ln), Lt $\geq$ 33, Lt $\leq$ 42, Ln $\geq$ -124.3, Ln $\leq$ -115,
ext:synonym(‘Manica’,‘bradleyi’,Ge$_1$,Sp$_1$), d$_4$(Ge$_1$,Sp$_1$,Ge$_2$,Sp$_2$),
ext:synonym(Ge$_2$,Sp$_2$,Ge,Sp).
q$_3$(d$_1$,Lt,Ln,Ge,Sp) :- d$_2$(Ge,Sp,Ct,Lt,Ln), Lt $\geq$ 33, Lt $\leq$ 42, Ln $\geq$ -124.3, Ln $\leq$ -115,
ext:synonym(‘Manica’,‘bradleyi’,Ge$_1$,Sp$_1$), d$_4$(Ge$_1$,Sp$_1$,Ge$_2$,Sp$_2$),
ext:synonym(Ge$_2$,Sp$_2$,Ge,Sp).
q$_3$(d$_1$,Lt,Ln,Ge,Sp) :- d$_3$(Mp,Cf,Lt,Ln), Lt $\geq$ 33, Lt $\leq$ 42, Ln $\geq$ -124.3, Ln $\leq$ -115,
ext:synonym(‘Manica’,‘bradleyi’,Ge$_1$,Sp$_1$), d$_4$(Ge$_1$,Sp$_1$,Ge$_2$,Sp$_2$),
ext:synonym(Ge$_2$,Sp$_2$,‘Manica’,‘parasitica’).
q$_3$(d$_1$,Lt,Ln,Ge,Sp) :- d$_3$(Mp,Cf,Lt,Ln), Lt $\geq$ 33, Lt $\leq$ 42, Ln $\geq$ -124.3, Ln $\leq$ -115,
ext:synonym(‘Manica’,‘bradleyi’,Ge$_1$,Sp$_1$), d$_4$(Ge$_1$,Sp$_1$,Ge$_2$,Sp$_2$),
ext:synonym(Ge$_2$,Sp$_2$,‘Aphaenogaster,‘calderoni).

4 Summary and Future Work

We have described an initial prototype that enables semantic-based dataset-discovery queries and supports mixing external services with traditional query-answering techniques. The prototype is written in Prolog and has an accompanying web interface for queries over geographic region, species, and parasitic relationship. We are extending the prototype by adding additional ontology-based query answering techniques including support for external services that perform transformation operations. To illustrate, the semantic description below is for a dataset similar to d$_1$, but uses an external service UTM2LatLon(Ux,Uy,Re,Zo,Lt,Ln) that converts UTM to latitude-longitude degree coordinates.

d$_5$(Ge,Sp,Co,Ux,Uy,Re,Zo) :-
Observation(O), value(O,Co), context(O,S), location(S,P), UTMPoint(P),
UTMx(P,Ux), UTMy(P,Uy), region(P,Re), zone(P,Zo), item(O,A), Abundance(A),
property(A,N), SciName(N), genus(N,Ge), species(N,Se).

To answer query q_1, we want to (1) return d$_5$ as a relevant source, since UTM points can be converted to latitude-longitude points using UTM2LatLon, and (2) correctly insert a call to UTM2LatLon into the resulting query as part of the query rewriting. We are currently exploring *parameter dependency* specifications for this purpose, in which the domain and range of an external service are semantically described. In general, we believe incorporating external services into mediator architectures provides a powerful framework to support complex integration and transformation of scientific data.

References

1. W. Berensohn. The concept of “Potential Taxa” in databases. *Taxon*, vol. 44, 1995.
2. S. Bowers and B. Ludäscher. An ontology-driven framework for data transformation in scientific workflows. In *Proc. of Data Integration in the Life Sciences*, LNCS, vol. 2994, 2004.

3. A. Y. Halevy. Answering queries using views: A survey. In *VLDB Journal*, 10(4), 2001.
4. B. Hölldobler and E. O. Wilson. *The Ants*. Harvard University Press, 1990.
5. N. F. Johnson. The Hymenoptera Name Server. http://atbi.biosci.ohio-state.edu:210/hymenoptera/nomenclator.home_page
6. T. Paterson and J. Kennedy. Approaches to storing and querying structural information in botanical specimen descriptions. To appear in *Proc. of BNCOD*, LNCS, July, 2004.
7. A. Purvis and A. Hector. Getting the measure of biodiversity. *Nature*, vol. 405, 2000.
8. A. Y. Levy, A. Rajaraman, and J. J. Ordille. Query-answering algorithms for information agents. In *Proc. of AAAI*, 1996.
9. B. Ludäscher, A. Gupta, and M. E. Martone. Model-based mediation with domain maps. In *Proc. of ICDE*, IEEE Computer Society, 2001.
10. D. R. B. Stockwell and D. P. Peters. The GARP modelling system: Problems and solutions to automated spatial prediction. *Intl. J. of Geographic Information Systems*, vol. 13, 1999.