# Finding Data, Knowledge, and Answers on the Semantic Web

Tim Finin, Joel Sachs, and Cynthia Sims Parr

Computer Science and Electrical Engineering
University of Maryland, Baltimore County
Baltimore, MD 21250 USA
{finin, jsachs}@cs.umbc.edu; csparr@cs.umd.edu

**Abstract**

Web search engines like Google have made us all smarter by providing ready access to the world's knowledge whenever we need to look up a fact, learn about a topic or evaluate opinions. The W3C's Semantic Web effort aims to make such knowledge more accessible to computer programs by publishing it in machine understandable form. As the volume of Semantic Web data grows, software agents will need their own search engines to help them find the relevant and trustworthy knowledge they need to perform their tasks. We will discuss the general issues underlying the indexing and retrieval of RDF-based information and describe Swoogle, a crawler based search engine whose index contains information on over two million RDF documents, and Tripleshop, which uses Swoogle to automatically build datasets appropriate for responding to user supplied queries. We will illustrate their use in ELVIS (Ecosystem Location Visualization and Information System), a distributed platform for constructing end-to-end use cases that demonstrate the semantic web's utility for integrating scientific data.

## Introduction

### How this experiment came to be

The data discovery, knowledge sharing, and collaboration problems faced by scientists are those the semantic web is designed to address (Hendler 2003, Finin and Sachs 2004, Zhao et al. 2004). The interdisciplinary areas of biodiversity and environmental biocomplexity, in particular, require collaboration and data sharing amongst specialists in the fields of ecology, conservation biology, and evolution, each of which has its own partially-shared vocabulary and way of seeing the world. We therefore identified this area as an excellent semantic web test bed.

The resulting SPIRE (Semantic Prototypes in Research Ecoinformatics) project[1] was funded three years ago by NSF to build prototypes exploring how the semantic web can address some of these problems. This paper describes two general-purpose tools that we have developed – Swoogle and Tripleshop – which, when taken together with our domain specific prototypes, enable experimentation with a large number of end-to-end semantic web use cases.

We proceed as follows: we conclude our introduction by giving background on invasive species. The next section describes Swoogle, a crawler based search engine whose index contains information on over two million RDF documents, and Tripleshop, which uses Swoogle to automatically build datasets appropriate for responding to user supplied queries. Sections three and four describe the sources of our data, namely ELVIS (the Ecosystem Location Visualization Information System), a suite of tools for predicting food webs, and a set of ontologies that we created to enable knowledge sharing. Section four also discusses some of the ontology engineering problems that we faced and continue to face. In section five, we discuss other approaches; in particular those based on the social web phenomenon, and speculate on how the social and semantic webs are likely to come together in the ecoinformatics domain.

### Background on Invasive Species

The specific domain of our use cases has been invasive species. Invasive species research is topical, depends on large numbers of distributed observations, and suffers from typical data integration problems. In this section, we give some background on the problem.

Species that are introduced into ecosystems in which they are not aboriginal are classified as *non-native* or *exotic*. Invasives are the small subset of non-native organisms that spread uncontrollably and therefore damage or displace native species, disrupt ecological processes and productivity, or threaten human health. Famous invasives include zebra mussels, the Asian longhorn beetle, Chinese snakehead fish. Not so famous invasives include sudden oak death, leafy spurge, and innumerable algae. Emerging diseases such as West Nile Virus can also be considered invasives. Several thousand weeds, crop pests, plant diseases, disease-vector insects, exotic predators, etc. are of active policy concern in the U.S. Invasive species are thought to be one of the two most important causes of declines and extinction of rare species, and cost the U.S. economy over $138 billion per year (Pimentel et al. 2000).

[1] http://spire.umbc.edu

Copyright © 2006, American Association for Artificial Intelligence (www.aaai.org). All rights reserved.

The invasive species problem is growing, as the number of pathways of invasion (ship ballast water, airplane wheel wells, highways, disease vectors, human agents, etc.) increases.

In general, once an invasive species has established itself in its new environment it is very difficult to eradicate; early detection is typically the key to a successful intervention. Thus, perhaps more than in any other discipline, the non-professional citizen scientist plays a vital role. The majority of new species invasions are first reported by amateurs, and reporting mechanisms have been established at the local, state, and national level. The semantic web, via tools such as the TripleShop described below, has the potential to tie these observations together with each other, and also to other data such as food web and natural history information.

## Finding Data and Answers

### Swoogle

Swoogle (Ding et al., 2004) is a crawler-based Semantic Web search engine that discovers and indexes documents containing RDF data. Running since 2004, it has indexed over two million such documents and has nearly 700 registered users. As new Semantic Wed Documents (SWDs) are discovered, Swoogle analyzes them to extract their data, compute metadata and derive statistical properties. The data is stored in a relational database and an information retrieval system (currently Lucene). In addition, a copy of the source document is stored and, since late 2005, added to an archive of all versions of every SWD discovered.

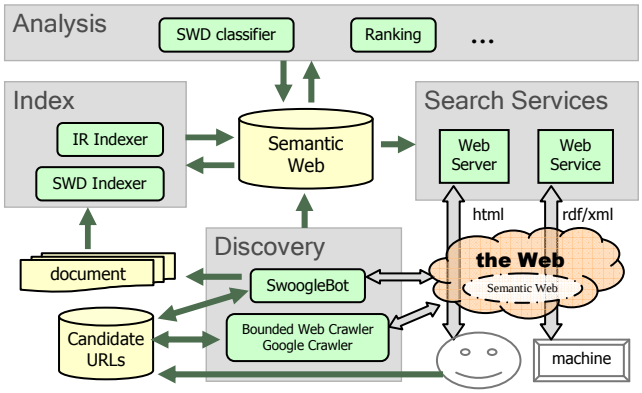


Figure 1. Swoogle uses an adaptive crawler to discover documents with RDF content. The RDF and document metadata are analyzed and stored in a database and indexed in an information retrieval system. Custom algorithms are used to rank ontologies, data documents and terms. Services are provided to both humans and programs.

Swoogle was designed with several use cases in mind. Semantic Web developers and researchers can use Swoogle to discover useful ontologies or terms and to collect data on properties of the Semantic Web (Ding and Finin 2006) and how it is being used (Ding et al. 2005). Semantic Web tools, such as TripleShop, use Swoogle through it's APIs to find data or ontologies for their users. Finally, software agents can use Swoogle to find RDF data and knowledge in support of performing their tasks.

**Discovery.** Swoogle currently uses Google to find initial seed documents that are likely to be SWDs. Other seeds come from user submissions. Since SWDs typically use special file extensions such as .rdf or .owl, Swoogle queries for files with such extensions. The extensions are dynamically selected (an extension is selected if more than 10 SWDs have used it and it has at least 50 percent accuracy in classifying SWDs). Since Google returns at most 1,000 results for any query, Swoogle takes advantage of its feature that restricts a search to results from a specified domain or site. Site queries work because of the locality hypothesis -- a Web site hosting more than two SWDs is likely to have more. Once it has discovered and validated an SWD, Swoogle uses a simple focused crawler to explore the Web environment around it to find more. After filtering out the non-SWDs from the results, Swoogle extracts a list of the sites on which the SWDs were found and uses them as seeds for further crawls as well.

**Ranking.** Google's success with its PageRank algorithm has demonstrated the importance of ordering the results that a query returns. Swoogle uses two custom ranking algorithms -- OntoRank and TermRank -- to order a collection of SWDs or terms, respectively. These algorithms are based on an abstract "surfing" model that captures how an agent might access Semantic Web knowledge. Navigational paths on the Semantic Web are defined by RDF triples as well as by the resource-SWD and SWD-SWD relations. However, revealing most of these connections requires a centralized analysis.

**Analysis**. Swoogle performs a very modest amount of reasoning using RDFS and OWL semantics due to the computational cost and potential for inconsistencies in dealing. However, it does compute many useful properties and still more can be easily derived from the information in its database. For example, Swoogle computes a document's ontology ratio as the fraction of its RDF triples that participate in defining a term. Documents with non-zero ontology ratios are considered to be ontologies in Swoogle's search services. Swoogle can also compute useful statistical measures, such as the conditional probabilities of two namespaces or two terms being used by a SWD.

### TripleShop

We first developed TripleShop as a component of Swoogle. It worked as follows: Swoogle would present query results (URIs) to the user, and then the user could check URIs to be added to his shopping cart. Eventually, a user could "check out", have all URIs loaded into Redland,

and be presented with an interface for issuing SPARQL queries. This utility proved to be an extremely useful tool in integrating scientific data[2] and so we implemented TripleShop as a stand alone service, with added functionality[3]. We describe this new functionality below.

## Current Features

**Finding Datasets** We added a "dataset finder" application that, in the absence of a FROM clause in the SPARQL query, searches Swoogle for URIs that contain terms contained in the WHERE clause. The user can then select which of these URIs she wants to query over, and also manually add URIs to the dataset.

**Constraints** A user might want to restrict her search for data in a number of ways. We allow constraints to be placed on the domain of a URI, and on namespaces that it uses. We will also soon enable all metadata that Swoogle has about a document to be the subject of constraints. This includes all assertions that a document makes about itself.

**Reasoning** After constructing a dataset, the user can specify a level of reasoning to be performed in executing the query. Choices range from no reasoning, through RDFS, to OWL.

**Dataset persistence** A user can save a dataset on the TripleShop server, tag a dataset, search for existing tagged datasets, and add tags to existing datasets. Datasets are stored as lists of URIs. A user can also choose to materialize a dataset, in which case the triples themselves are stored in a database.

We envision a scenario where a user begins by issuing a few illustrative queries (with no FROM clause!). TripleShop then gathers and indexes all triples that might be relevant to the query, perhaps also forward chaining to generate all implied triples. This process may take anywhere from seconds to hours. When it's complete, the user can query against the resulting datastore, and can tag it appropriately for other users to find.

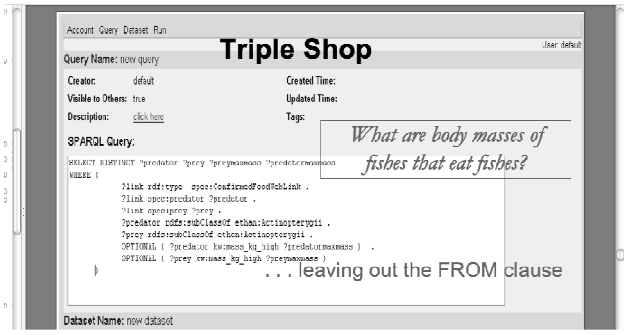


**Figure 2.** "Show max body masses and feeding data for all fish-eating fish." is one of several stored queries tagged "spire". This query implicitly defines a dataset, namely all URIs considered by Swoogle to be potentially relevant to the query.

[2] http://spire.umbc.edu/ont/sparql_demo/query.php?demo=1

[3] Accessible at http://sparql.cs.umbc.edu/tripleshop2; contact authors for a login account.

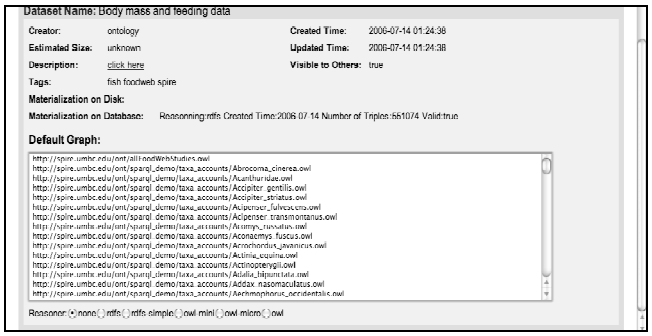


**Figure 3.** A stored dataset comprising URIS containing (according to Swoogle) body mass or feeding information for fish.

## Eco-Resources in OWL

The utility of Swoogle and Tripleshop derives from the data available to them. Although we anticipate an impending avalanche of RDF (see "Other Approaches to Web Semantics", below), currently, semantics must be squeezed from the web like juice from a dried-out lemon. In addition to Swoogle and Tripleshop, our main prototype is ELVIS – a suite of tools for food web prediction. At the beginning of SPIRE, we pledged that both our input and output data would be RDF. And so the components of ELVIS have become our main source of data for Tripleshop queries. Specifically, our RDF data primarily consists of:

i. species distribution data compiled by the California node of the National Biological Information Infrastructure;
ii. trophic data compiled from over 250 datasets;
iii. the complete contents of Animal Diversity Web (ADW), a popular on-line encyclopedia (Myers et al. 2007); and
iv. a collection of lists designating species as being invasive in particular regions.

We describe this data, and how it came to be on the semantic web, below.

## Motivation

ELVIS is motivated by the belief that food web structure plays a role in the success or failure of potential species invasions. Because very few ecosystems have been the subject of empirical food web studies, response teams are typically unable to get quick answers to questions like "what are likely prey and predator species of the invader in the new environment?" The ELVIS tools seek to fill this gap.

ELVIS expresses all data in OWL via a collection of ecological and evolutionary ontologies. This, together with our service-oriented architecture, enables much flexibility in integrating with other semantic web applications.

The task of providing food web information for a user-specified location breaks into two distinct problems: constructing a species list for a given location; and constructing a food web from a given species list (and habitat information). We describe each in turn.

## The Species List Constructor

Our goal is to allow a user to input a location, and get back a species list for that location. This is a hard problem, typically *ad hoc*, and relying on expert knowledge. In California we currently use: (i) park inventories; (ii) point locations, e.g. from specimen descriptions in museums and herbariums; and (iii) distribution maps generated by applying statistical techniques to point locations. The ontologies and synthesis strategies we have developed should apply to other states.

In support of the effort to return species lists for particular locations, CAIN (the California Invasive species Node of the National Biological Information Infrastructure, and a SPIRE partner) has created two web services on the CAIN server. In the first of these, CAIN provides a list of the terrestrial vertebrates in the state to the county scale, using the California Wildlife Habitat Relationships (CWHR) database. This database provides life history details for the terrestrial vertebrates (mammal, reptiles, amphibians, and birds) of California, including information on habitat and geographic range. CAIN extracted the range information by county for this database, converted it into RDF, and placed it into a Kowari RDF data store queryable using a SOAP interface.

The second web service resides on top of CAIN's CRISIS Maps application (CRISIS) for displaying observations of invasive weeds in California and the Southwest, and uses an OpenGIS Web Feature Service (WFS) (OGC 2005). WFS is a protocol that allows clients to retrieve and update geospatial data encoded as vector features over the Internet. This service returns point observations of selected weed species within a latitude/longitude bounding box in Geographic Markup Language (GML) (OGC 2004).

Species List Constructor data is exposed in OWL via the CWHR ontology. California Wildlife Habitat Relationships (CWHR) is an information system run by California's Department of Fish and Game. It contains life history, geographic range, habitat relationships, and management information on 692 species of amphibians, reptiles, birds, and mammals known to occur in the state. The CWHR ontology expresses all this information in OWL, and is our main means of expressing data for the species list constructor.

## Food Web Constructor

The Food Web Constructor (FWC) uses empirically known food web links from multiple sources to predict links among a list of focus organisms (taxa) of interest to a user. Our current algorithm uses taxonomic distance to weight evidence supporting or failing to support links among the focus taxa. Each suspected link is reported, together with references to supporting evidence. Summary statistics of the resulting food web, such as number of predicted links and connectivity, are also reported.

Our goal is to make FWC a platform for experimenting with different approaches to food web prediction. Currently, a user can set different parameters and weights for the prediction algorithm. In the future, we can use semantic web tools too populate the training database. We would like to provide users with the ability to choose amongst prediction algorithms, or to provide their own (as a web service). We already provide a mechanism to assess the success rate of the different algorithms or model parameters, and report such statistics as accuracy, precision, and recall.

## Evidence Provider

As the computer scientists on our team have become more familiar with the ecological issues involved, our thinking of what the semantic web can/should contribute to invasive species science has matured. The massive uncertainty in so many areas of ecology has led us away from thinking of our applications as 'answer providers', and towards thinking of them as 'evidence providers'. This is reflected in our Evidence Provider (EP) tool.

Given a list of n species, there are n squared possible trophic links. The Evidence Provider allows a user to drill down on a potential link to see the evidence for and against it. This includes actual observed links, the study in which they were published, and the relationship between the species in the observed link and the predicted link.

FWC and EC input and output are expressed in RDF via the SpireEcoConcepts ontology. This ontology defines the terms necessary to i) express both confirmed and potential trophic links; ii) describe bibliographic information of food web studies; iii) provide ecosystem labels (montane, riparian, etc.); and iv) represent taxonomic ranks and distances.

## ETHAN

ETHAN (the Evolutionary Trees and Natural History ontology) arose out of our need to represent taxonomic, phylogenetic, and natural history information in OWL (Parr et al. 2006). We do this via two core OWL-DL ontologies. First, several hundred thousand scientific names of species and higher taxonomic levels are represented in a class hierarchy, without biological ranks. These data come from ITIS, the Integrated Taxonomic Information System, and from a number of smaller phylogentic trees. An online utility at http://spire.umbc.edu/ allows a user to generate parts of the ontology of interest to their own work. Second, an ETHAN keyword ontology organizes natural history concepts, such as reproductive and physical description categories, as well as quantitative measures such as body mass

and lifespans. This natural history information comes from the Animal Diversity Web (ADW). Although there are several "species page" web sites, we chose to ontologize ADW first, since members of our team were formerly involved in ADW development, and were able to secure the cooperation of the ADW technical lead. All ADW species accounts are now available as OWL documents, and publishing in OWL is a part of the weekly ADW publishing process. We believe that this example will help to persuade other species banks (such as Fishbase[4]) to follow our lead, and to publish their data on the semantic web.

In composing ETHAN, we made a number of modeling decisions. We describe one of the most controversial of them below.

**Property or subclass?**

Suppose we want to say that lions are carnivores. It seems natural to say:

(1) <Panthera_leo> <subClassOf> <Carnivora>

This is what we do in ETHAN. What if we want to say lions live in the savanna? It seems natural to say that:

(2) <Panthera_leo> <hasHabitat> <savanna>,

where, perhaps, the hasHabitat property is elsewhere defined as having a range which includes either the savanna class, or the literal "savanna".

We don't do this in ETHAN. Instead, we say:

(3) <Lions> <subClassOf> <SavannaLivingThing>

More generally, we find that the semantics behind an arbitrary relation can often be expressed using <subClassOf> relations. Doing so has the following benefits:

1. This seems to be more computationally efficient. (We have no hard evidence for this.)
2. This makes it easy to introduce a new concept, especially in a distributed manner. This is especially true if the concept corresponds to a list, as is often the case in ecology. For example, a species can be listed (by any of a number of organizations) as threatened or endangered, or invasive in a particular area. If a treaty, NewTreaty, lists a number of species to be considered invasive, we represent this by introducing a <NewTreatyThing> class, and making the appropriate subClass assertions.
3. This leads to fewer disagreements among scientists and, therefore, greater chance of ontological adoption. We are, essentially, equating classes with sets.

By adding

(4)<SavannaLivingThing> <subClassOf> <EarthLivingThing>

we are able to query for the habitat of lions by querying for all classes that are both superclasses of <Panthera_leo> and also subClasses of <EarthLivingThing>

This approach, of course, results in an abundance of multiple inheritance. Some people say this is a bad thing, on the grounds that it is undisciplined and ad hoc.

### Integrating Food Web and Natural History Data

We have been using TripleShop to integrate food web data, taxonomic information, and natural history data (Sachs et al. 2006). For example, Figure 2 shows a query that combines data from two ontologies – taxonomic and natural history information from ETHAN and food web data from SpireEcoConcepts – and from the ELVIS database to retrieve body masses of fish-eating fish. Figure 3 shows the datasets returned by Swoogle as being potentially relevant. Since most ecological analysis is done with statistical or spreadsheet software, users can choose to get the results back as CSV or Excel files, in addition to the standard HTML and XML representations.

## Other Approaches to Web Semantics

The role of RDF and OWL in integrating web resources has been somewhat usurped by web 2.0 technologies such as microformats (Khare 2006). For example, five years ago, we envisioned mashups as one of the capabilities that the semantic web would enable. As it turns out, XML, XHTML, and open web service APIs are sufficient for most mashups. When RDF does come into play, it is usually in the form of RSS feeds.

However, some applications will benefit from the richer semantics made possible by OWL representations. For example, we have begun experimenting with reporting observational data (e.g. species X has been observed in location Y) in a variety of lightweight formats, (RDF/a (Adida and Birbeck 2006), microformats, and structured blogging). Simply visualizing such observations on a map is not difficult and we have begun to do this using Splickr, a mash-up of Spire and Flickr.[5]

A surveillance program may expect a vast amount of observational data, much of which is not relevant. In order for intelligent agents to successfully filter out only the observations of interest to invasive species biologists, richer representations of related information are necessary. For example, according to various sources, species X has already been reported as invasive in area Y, so this observation probably not interesting to a biologist. Or, species X is related to another species that has had high success invading related habitats in other regions, so perhaps it is of great interest. The information needed to make these inferences rests on complex and inconsistent vocabularies. It will also be dynamic and highly distributed. Therefore it makes sense to provide it in OWL ontologies so that tools

[4] http://www.fishbase.org

[5] http://spire.umbc.edu/splickr/

like TripleShop and Swoogle can harvest, integrate, and query it automatically.

ETHAN provides the ontological framework for publishing invasive species lists (resources provided by numerous governmental agencies) and the taxonomic and habitat information needed to interpret such lists. More lightweight formats may be better to support the observations themselves.

## Future Work

All of the prototypes described above remain in development. In TripleShop we currently handle conflicts amongst sources by ensuring that they don't occur. Obviously, this approach will not scale. We may add to TripleShop a quarantine area for triples that conflict with the current graph or each other. The user could then choose which to include in the dataset. It is likely that contradictory triples will surface only late in the process, after reasoning is applied, and some experimentation will be required to determine the optimal placement of the quarantine in the workflow.

We would also like to put TripleShop at the service of analytical tools wishing to populate local databases, such as the Food Web Constructor. In the future we will add a notification service to TripleShop to alert a user as soon as new data matching a query becomes available on the semantic web. Finally, we plan improvements to the user interface, performance tuning, and, possibly, experimentation with various approaches to parallelization.

For Fieldmarking we plan to formalize our experimentation into the integration of lightweight observational data reporting with more heavyweight ontologies. First we will determine the advantages and disadvantages of several lightweight formats and their ability to be harvested and integrated by tools such as TripleShop. Then we will develop a GreaseMonkey script to allow users to easily generate such machine-readable data in blogs or discussions.

## Acknowledgements

This research was supported by NSF ITR 0326460 and matching funds received from USGS National Biological Information Infrastructure. In addition to cited collaborators, we thank Roger Espinosa

## References

Adida, B. and Birbeck, M. 2006. RDFa Primer 1.0, W3C Working Draft, http://www.w3.org/TR/xhtml-rdfa-primer/.

CRISIS Maps http://cain.nbii.gov/cgibin/mapserv?map=../html/cain/crisis/crisismaps/crisis.map&mode=browse&layer=state&layer=county

Ding, L.; Finin, T.; Joshi, A.; Pan, R.; Cost, R. S.; Peng Y.; Reddivari, P.; Doshi, V. C.; Sachs, J. 2004. "Swoogle: A Search and Metadata Engine for the Semantic Web", In Proceedings of the Thirteenth ACM Conference on Information and Knowledge Management, Washington, D.C.: ACM.

Ding, L.; Pan R.; Finin, T.; Joshi, A.; Peng Y.; Kolari, P. 2005. Finding and Ranking Knowledge on the Semantic Web, In *Proceedings of the 4th International Semantic Web Conference*, pp. 156-170. Galway, Ireland:Springer-Verlag.

Ding, L. and Finin, T. 2006. Characterizing the Semantic Web on the Web, In *Proceedings of the 5th International Semantic Web Conference*. Athens, GA.: Springer Verlag.

Finin, T. and Sachs, J. 2004. Will the Semantic Web Change Science?, *Science Next Wave*, September 2004.

Hendler, J. 2003. Science and the semantic web, *Science*. Volume 299, no. 5606, 520-521.

Khare, R. 2006. Microformats: the next (small) thing on the semantic Web?, *IEEE Internet Computing*, volume 10, no. 1, 68-75.

Myers, P.; Espinosa, R.; Parr, C. S.; Jones, T.; Hammond, G. S.; and Dewey, T. A. 2006. Animal Diversity Web. http://animaldiversity.org.

OGC (Open Geospatial Consortium), Inc. 2004. OpenGIS Geographic Markup Language (GML) Encoding Specification Version 3.1.1.

OGC (Open Geospatial Consortium), Inc. 2005. Web Feature Service Implementation Specification, Version 1.1.0.

Pimentel, D.; Lach, L.;Zuniga, R. ; and Morrison,D. 2000 Environmental and economic costs associated with non-indigenous species in the United States. *Bioscience* 50:53-65.

Parr, C. S.; Sachs, J.; Parafiynyk, A.; Wang, T.; Espinosa, R.; Finin, T. 2006. ETHAN: The Evolutionary Trees and Natural History Ontology, Technical Report, Dept. of Computer Science and Electrical Engineering, UMBC.

Sachs, J.; Parr, C. S.; Parafiynyk, A.; Pan, R.; Han, L.; Ding, L.; Finin, T.; Hollander, A.; and Wang, T. 2006. Using the Semantic Web to Support Ecoinformatics, In Proceedings of the AAAI Fall Symposium on the Semantic Web for Collaborative Knowledge Acquisition. Washington D.C.: AAAI.

Zhao, J.; Wroe, C.; Goble, C.; Stevens, R.; Quan, D.; Greenwood, M. 2004.Using Semantic Web Technologies for Representing E-science Provenance, *In Proceedings of the 3rd International Semantic Web Conference*, 92-106. Sanibel Island, FL.: Springer-Verlag.