Using the Semantic Web to Support Ecoinformatics

Joel Sachs jsachs@umbc.edu
Cynthia Sims Parr csparr@umbc.edu
Andriy Parafiynyk andr1@umbc.edu
Rong Pan pan.rong@umbc.edu
Lushan Han lushan1@umbc.edu
Li Ding ding.li@umbc.edu
Tim Finin finin@umbc.edu

Dept. of Computer Science and Electrical Engineering
University of Maryland Baltimore County
Baltimore, MD 21250 USA

Allan Hollander
Information Center for the Environment
University of California Davis
Davis, CA 95616

David Wang
Dept. of Computer Science
University of Maryland
College Park, MD, 20742 USA
Tw7@cs.umd.edu

Abstract

We describe our on-going work in using the semantic web in support of ecological informatics, and demonstrate a distributed platform for constructing end-to-end use cases. Specifically, we describe ELVIS (the Ecosystem Location Visualization and Information System), a suite of tools for constructing food webs for a given location, and Triple Shop, a SPARQL query interface which allows scientists to semi-automatically construct distributed datasets relevant to the queries they want to ask. ELVIS functionality is exposed as a collection of web services, and all input and output data is expressed in OWL, thereby enabling its integration with Triple Shop and other semantic web resources.

1. Introduction

The data discovery, knowledge sharing, and collaboration problems faced by scientists are often the types of problems that the semantic web is meant to address, and scientists have been amongst the first adopters and co-creators of the semantic web. This is especially true in the highly interdisciplinary area of environmental biocomplexity, an area requiring collaboration and data sharing amongst specialists in the fields of systematics, ecology, and evolution, each of which has its own partially shared vocabulary and way of seeing the world.

The SPIRE (Semantic Prototypes in Research Ecoinformatics) project [13] 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 of our prototypes, which, when taken together, enable experimentation with a large number of end-to-end semantic web use cases. The domain of our investigation has been invasive species, due to its topicality, its dependance on large numbers of distributed observations, and its inheritance of the problems mentioned above that are ever-present in biocomplexity research.

The paper proceeds as follows: we conclude our introduction by giving background on invasive species, and describing related work. In section 2, we describe ELVIS (the Ecosystem Location Visualization Information System), a suite of tools for predicting food webs. Section 3 describes the ontologies that we created to enable knowledge sharing, and discusses some of the problems that we faced and continue to face. Section 4 presents Triple Shop, a tool which allows a user to specify arbitrary SPARQL queries, with or without a FROM clause, together with a reasoning level. Thus, we are able to integrate and reason over diverse biocomplexity data in response to ad-hoc queries. We conclude with a discussion of future work.

1.1 Background on Invasive Species

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, through uncontrolled spreading, damage or displace native species, disrupt ecological processes and productivity, or threaten human health. Famous invasives include zebra mussels, the asian longhorn beetle, west nile virus, and chinese snakehead fish; not so famous invasives include sudden oak death, leafy spurge, and innumerable algae. 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 [12]. 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 Triple Shop 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.

1.2 Related Work

Previous work on data integration in ecological informatics includes online data repositories [4] and workflow [9] ontologies. Individual food web researchers maintain and share their own digital data archives, in individualized data formats, though more accessible standardized archives are beginning to emerge [6]. There are good databases on invasive species (e.g. http://www.issg.org/) but they are not automatically integrated with information about non-invasive species with which they interact. To our knowledge, there does not exist web-based support for modeling an invasive species.

Our Tripleshop builds heavily on the Joseki SPARQLer service [7]. Our contribution has been to introduce features such as reasoning capabilities, the ability to automatically construct datasets relevant to a query, and the ability to store and tag datasets.

2. ELVIS

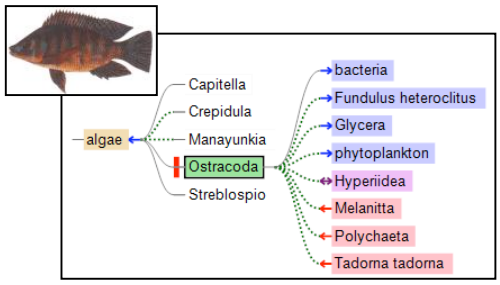

Figure 1. Nile Tilapia, an invader in Florida ecosystems, is predicted by ELVIS to eat algae and have no potential predators. Organisms predicted to be nearby in the food web (to the right of algae) could be impacted by or mediate the introduction of this competitor.

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.

2.1 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. There are, in general, three kinds of information that can be used to generate a species list: (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. We are integrating all of the above for California, and expect that the ontologies and synthesis strategies we have developed will apply to other states, and enable ELVIS to spread beyond California.

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 [3] for displaying observations of invasive weeds in California and the Southwest, and uses an OpenGIS Web Feature Service (WFS) [10] interface. (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) [11].

2.2 The 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

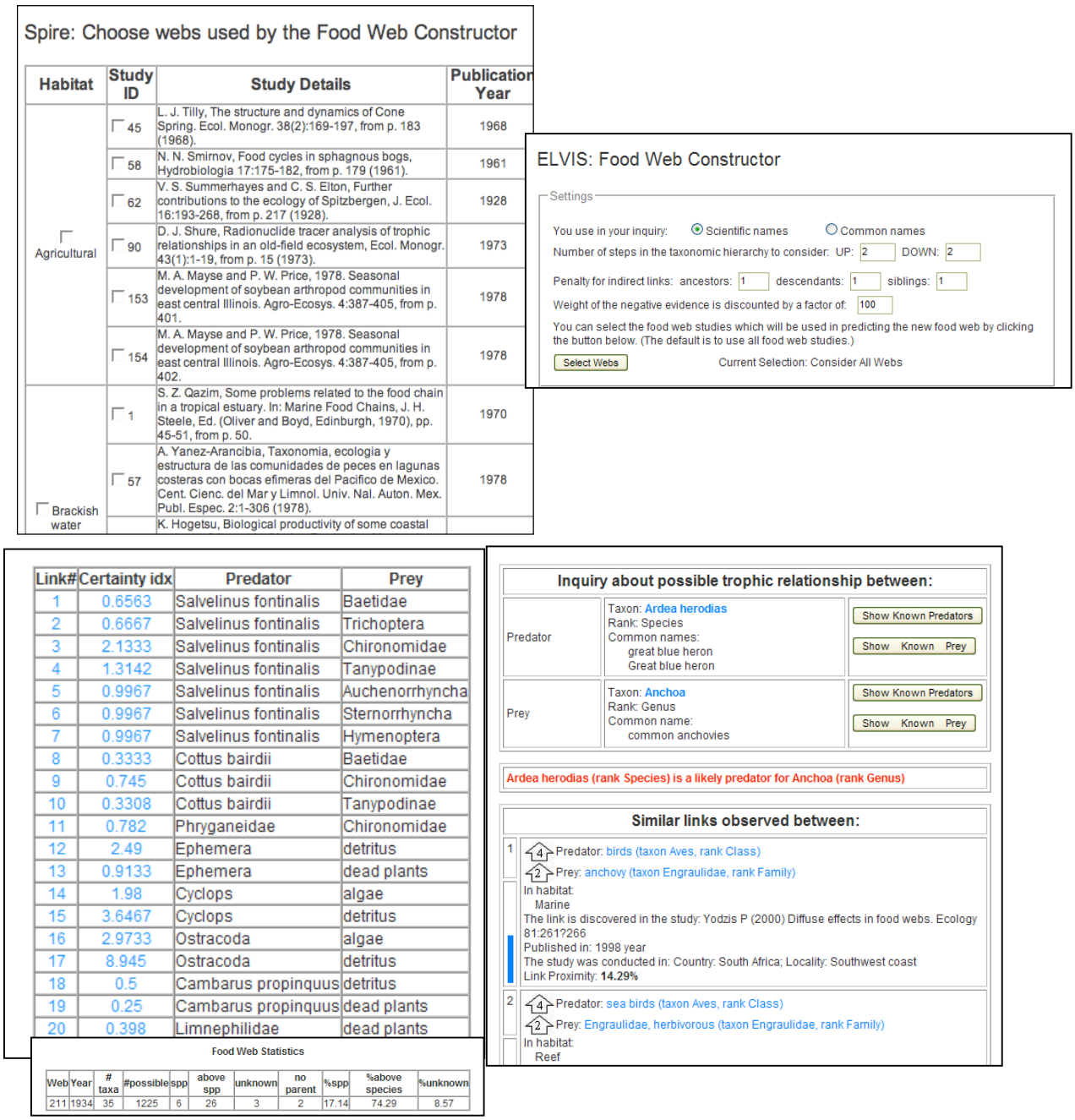

Figure 2. Elvis screen shots . Food web studies can be selected individually or by habitat (upper left). Different parameters and weights (upper right) can be used to compute the certainty index (the sum of the weights of positive and negative links). Each suspected link is reported, with the certainty index. Summary statistics of the food web are also reported (bottom left). Link# and Certainty index link to more detailed information and supporting evidence, as shown in the evidence provider view (lower right).

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. Food Web Constructor provides a number of ways to customize the analysis. A user can choose which individual food web studies to use or exclude from 257 datasets we compiled from previously digitized literature [described in Parr, in prep]. Studies can also be selected in groups by habitat, as it may be biologically reasonable to include only links from specific terrestrial or aquatic systems when making predictions. Focus taxa can be entered as simple text lists of common or scientific names, or XML files with either ITIS TSN's or scientific names.

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 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. (see Figure 2.)

2.3 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 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. (See Figure 2.)

3. Ontologies

3.1 ETHAN

ETHAN (the Evolutionary Trees and Natural History ontology) arose out of our need to represent taxonomic, phylogenetic, and natural history information in OWL. 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) [1]. 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. With ETHAN development nearing completion, all ADW species accounts will soon be available as OWL documents, and publishing in OWL will become a part of the weekly ADW publishing process. We believe that this example will help to persuade other species banks (such as Fishbase [5]) to follow our lead, and to publish their data on the semantic web.

In composing ETHAN, we faced a number of modeling problems, some of which are not yet resolved. For example, we currently express the fact that ginko-toothed beaked whales live in the Indian Ocean as follows:

```
<owl:Class rdf:ID="Mesoplodon_ginkgodens">
<kw:geographic_range>
rdf:datatype="http://www.w3.org/2001/XMLSchem
a#string">Indian ocean
</owl:Class>
```

This enables us to issue simple sparql queries to retrieve all species that live in the Indian Ocean. However, OWL has no provision to map assertions about a class to members of the class, so we haven't said anything about where individual whales live. Thus, there is a sense in which what we really mean is

```
<owl:Class rdf:ID="Mesoplodon_ginkgodens">
<rdfs:subClassOf>
<owl:Restriction>
<owl:onProperty
rdf:resource="kw:geographic_range"/>
<owl:hasValue>indian ocean</owl:hasValue>
</owl:Restriction>
</rdfs:subClassOf>
</owl:Class>
```

However, sparql provides no mechanism to retrieve all classes with a particular owl:Restriction, so the above formulation leaves us unable to query on species that live in the Indian Ocean.

Since our use cases all deal with species (or, more generally, taxa), and not individual organisms, we are having no problems with the first formulation. However, we have received requests from other research groups interested in using our ontologies, and we plan on adjusting ETHAN to accommodate more general use cases.

3.2 SpireEcoConcepts

We developed SpireEcoConcepts to enable describing ELVIS input and output in OWL. The 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.

3.3 CWHR

CWHR (California Wildlife Habitat Relationships) is an information system run by California's Dept. 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.

4. The SPIRE Triple Shop

4.1 Evolution of Triple Shop

We first developed Triple Shop as a component of our Swoogle semantic web search engine [8]. Swoogle crawls the semantic web, computing and storing metadata for each page, including 'Ontology Rank' (our semantic web version of Google's PageRank), an estimate of how important a page is to the semantic web. Swoogle currently indexes more that 1.5 million semantic web documents supported by about ten thousand ontologies.

Triple Shop originally 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 (see, e.g. http://spire.umbc.edu/ont/sparql_demo/query.php?demo=), and so we implemented Triple Shop as a stand alone service, with added functionality (http://sparql.cs.umbc.edu/tripleshop2 - contact authors for a login account). We describe this new functionality below.

4.2 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 URI's 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 Triple Shop 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.

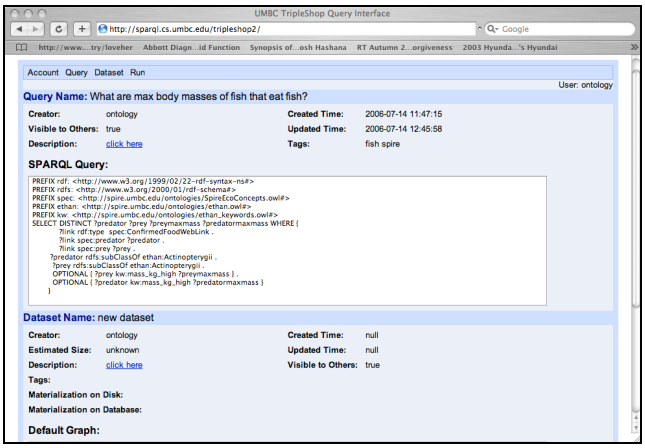

Figure 3. "Show max body masses and feeding data for all fish-eating fish." This is one of several stored queries tagged "spire". This query implicitly defines a dataset, namely all URIs considered by Swoogle to be potentially relevant to the query.

We envision a scenario where a user begins by issuing a few illustrative queries (with no FROM clause!). Triple Shop 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.

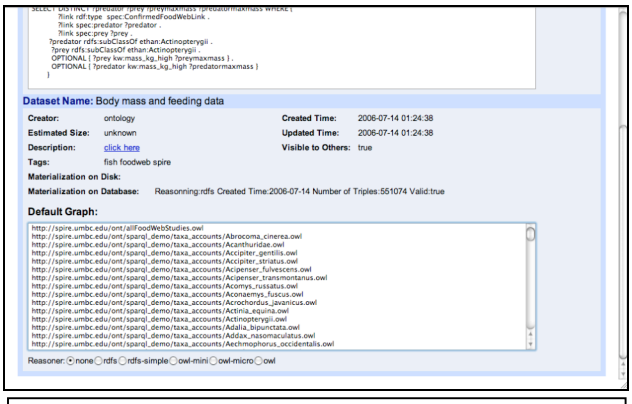

Figure 4. A stored dataset comprising URIS containing (according to Swoogle) body mass or feeding information for fish.

4.3 Using the Triple Shop to Integrate Food Web and Natural History Data

We have been using Triple Shop to integrate food web data, taxonomic information, and natural history data. For example, Figure 3 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 5 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.

5. Future Work

All of the prototypes described above remain in development. Future work includes:

Species List Constructor. We plan to move beyond the queries described in section 2.1, (i.e. "show me the species list for ____") to other, more subtle, queries that will illustrate the power of the semantic web more clearly. For example, a user will be able to select a species, and see the provenance of the claim that that species is present in a particular location; or the user will be able to simply query (e.g.): 'what is the evidence for and against the presence of bobcats in my backyard?'.

Food Web Constructor. The heart of the Food Web Constructor is our food web prediction algorithm, and we are engaged in experiments to determine optimum parameters for the algorithm. As mentioned above, we also plan on experimenting with entirely different models.

Triple Shop. We currently handle conflicts amongst sources by ensuring that they don't occur. Obviously, this approach will not scale. We may add to Triple Shop 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 Triple Shop at the service of analytical tools wishing to poulate 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.

6. Acknowledgements

This research was supported by NSF ITR 0326460 and matching funds received from USGS Nat. Bio. Information Infrastructure.

7. References

[1] Animal Diversity Web http://animaldiversity.ummz.umich.edu/site/index.html

[2] D. Brickley and R.V. Guha, RDF Vocabulary Description Language 1.0: RDF Schema, Feb. 2004; http://www.w3.org/TR/rdf-schema/.

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

[4] Dunne, J. A. The network structure of food webs. In: Ecological Networks: Linking Structure to Dynamics in Food Webs, eds. Pascual, M. and Dunne, J. A. Oxford University Press, 2005.pp. 27-86.

[5] Fishbase http://www.fishbase.org

[6] Jones , M. B.; Berkley, C.; Bojilova, J.; Schildhauer, M. P. 2001. Managing scientific metadata. IEEE Internet Computing. Vol: 5(5). Pages 59-68.

[7] Joseki http://www.sparql.org/query.html

[8] Li Ding *et al*., "Swoogle: A Search and Metadata Engine for the Semantic Web", Proceedings of the Thirteenth ACM Conference on Information and Knowledge Management, November 2004.

[9] Ludaescher, B. et al. 2004. Scientific workflow management and the Kepler system. Concurrency and Computation: Practice and Experience.

[10] Open Geospatial Consortium, Inc. 2005. Web Feature Service Implementation Specification, Version 1.1.0.

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

[12] Pimental et al. 2000 Environmental and economic costs associated with non-indigenous species in the United States. Bioscience 50:53-65.

[13] Semantic Prototypes in Research Ecoinformatics http://spire.umbc.edu