Improving Data Discovery for Metadata Repositories through Semantic Search

Chad Berkley[1], Shawn Bowers[2], Matthew B. Jones[1], Joshua S. Madin[3], Mark Schildhauer[1]
[1]National Center for Ecological Analysis and Synthesis, UC-Santa Barbara;
[2]Univ. of California, Davis; [3]Department of Biological Sciences, Macquarie University
berkley@nceas.ucsb.edu, sbowers@ucdavis.edu, jones@nceas.ucsb.edu,
jmadin@bio.ma.edu.au, schild@nceas.ucsb.edu

Abstract

The amount of ecological data available electronically is increasing at a rapid rate, e.g., over 15,000 data sets are available today in the Knowledge Network for Biocomplexity (KNB) alone. Using the existing search capabilities of these online data repositories, however, scientists struggle to quickly locate data that are relevant to their needs or that will integrate with their current data sets. Semantic technologies aim at addressing many of these problems and hold the promise of enabling more powerful "smart" searches of online data archives. We describe new semantic search features within the Metacat metadata system, which is used by many ecological research sites around the world for archiving their data using a standardized metadata format. Our semantic search system adds to Metacat the ability to store OWL-DL ontologies in addition to semantic annotations that link data set attributes to ontology terms. Our approach also extends Metacat to improve metadata search in multiple ways: (i) by expanding standard keyword searches with ontology term hierarchies; (ii) by allowing keyword searches to be applied to annotations in addition to traditional metadata; and (iii) by allowing more structured searches over annotations via ontology terms. We describe our implementation of these extensions, and compare and contrast these different types of search for a corpus of annotated documents. As data repositories continue to grow, these tools will be instrumental in helping scientists precisely locate and then interpret data for their research needs.

1. Introduction

A wide variety of data are used in ecological and environmental studies. Data for these studies quantify, among other things, the distribution and abundance of organisms; the processes that influence biological populations, communities, and ecosystems; and the environmental and anthropogenic drivers of these processes. Scientists increasingly rely on accessing and analyzing data collected by cross-disciplinary communities of researchers to achieve synthetic, crosscutting insights into the environment that can address issues of fundamental importance to science and society [1-3].

Data repositories can play an important role in increasing the frequency, scope, and efficiency of these synthetic studies. However, to be useful in such studies, data must be easy to discover and readily available and accessible. While some improvements have been seen, both of these issues—depth of holdings and effective data discovery—are still problematic for most repositories.

The holdings of data repositories in ecology have been growing rapidly [4]. For example, the Knowledge Network for Biocomplexity (KNB) data repository has grown exponentially to now contain over 15,000 data sets [5]. These archives hold tremendous promise for increasing the scope and efficiency of synthetic studies by lowering the barriers to finding and utilizing the broadest set of appropriate data for analysis. Nevertheless, these archives have far to go before they will represent a reasonable portion of the ecological, environmental, and related data that are collected each year.

Even at current collection sizes, however, the precision and recall of data searches in many repositories is not satisfactory. We use the standard definitions here for precision and recall: *precision* representing the proportion of relevant results out of all results returned; and *recall* being the proportion of relevant items found out of the total of relevant items available [6]. Data archives like the KNB, the National Biological Information Infrastructure (NBII) Metadata Clearinghouse, and the Global Change Master Directory (GCMD) rely on semi-structured metadata with fields containing largely natural-language descriptions to provide search and browsing capabilities and to allow human use and interpretation of the data. These metadata enable simple keyword searches that return results generally related to the topics of interest, but they cannot be used to perform precise searches of the data archives. For example, a search for the keyword 'soil' returns over 2000 data sets from the KNB, many of which are not data about soil *per se,* but rather have metadata documents that address the soil characteristics of the site at which data were collected. Thus, ironically data sets with more extensive (natural language) metadata are included in search results simply due to the incidental mention of a term in an ancillary part of the metadata document. These extraneous results decrease the precision of the search, seriously reducing the efficiency in researchers' finding the data they need.

Because natural-language metadata does not generally rely on controlled vocabularies, researchers typically classify their data sets using an *ad-hoc* set of descriptive terms. This in turn leads to issues with recall. Given the

number of synonymous, polysemous, and overlapping terms used in scientific disciplines, searches frequently miss relevant data because the search terms do not syntactically match the terms used in classifying the documents. While relatively simple to implement, string-based searches cannot provide the type of recall or precision needed by scientists trying to find useful data.

Libraries have long been concerned with providing more effective mechanisms for information retrieval, and this need has compounded over the last few decades by the explosion of available digital data [7]. These efforts have motivated research into the development of online search systems based on controlled vocabularies and thesauri, and comparative analyses of these systems relative to full text indexing and other natural language methods [8]. Even when controlled terms are used, the broad range of data needed by environmental scientists makes searches susceptible to "semantic drift", due to varying usage of terms within different disciplines.

More recently advances in algorithms used by Web search engines such as Google's PageRank have enabled powerful information retrieval from extremely large, distributed repositories of connected digital information [9]. Still, these approaches do not effectively bridge the gap between the retrieval of (e.g., Web) documents, and the scientists' need for precise discovery and interpretation of research data sets. The progression towards concept-based searches of Web-based information (e.g., [10-12]), however, represents a promising mechanism for addressing the needs for precise and potentially deep descriptions of data.

Standardized approaches for describing Web metadata are rapidly advancing, and frameworks are emerging for developing rich, semantic searches that can closely map to the structures and content common in scientific research data [13].

In this paper, we describe search approaches that exploit the use of formal reasoning over an ontology designed to facilitate the semantic description of scientific observations [14]. Specifically, the *Extensible Observation Ontology* (OBOE) provides a high-level abstraction of scientific observations and measurements that facilitate the creation of domain-specific vocabularies for defining observation and measurement semantics [15]. OBOE is represented using OWL-DL [16] and enables data (or metadata) structures to be linked to domain-specific ontology concepts so that critical aspects of scientific observations can be documented—such as what *kind of Entity* was measured, which *Characteristics* of that entity were measured and by which *Measurement Standards* (e.g., kilograms/m^2), and what other observations provide *Context* for interpreting measurements [15,17]. In our approach, *semantic annotations* are used to map these critical parts of a scientific observation to the data instances that are stored in a data repository.

In addition to plain-text keyword search, we describe and compare three different search methodologies to investigate the utility of semantic methods for scientific data discovery: (i) simple term expansion against ontologies to broaden the search terms against the metadata corpus; (ii) term expansion against semantic annotations; and (iii) structured searches that pose queries against the components of an observation described via OBOE.

The rest of this paper is organized as follows. Section 2 describes our metadata, ontology, and semantic annotation framework. Section 3 describes the different semantic-search techniques discussed above, and our prototype implementation of these based on the Metacat metadata management system (employed by the KNB). Section 4 concludes by summarizing our results and describing future work.

2. The Framework

Figure 1 shows the primary components of our semantic-discovery framework. The bottom of Figure 1 consists of two simple, example data sets. Although different types of data are often used within ecological analysis (e.g., raster, GIS, etc.), data sets are predominantly tabular (relational) and denote sets of related observations and measurements that were either directly collected or were the result of aggregation or analysis. Although not obvious, the example data sets in Figure 1 contain largely similar information consisting of spatial locations divided into sub-locations (i.e., a plot or quadrat), fertilization treatment information, and weight measurements.

Ecological data is often stored within simple text files or spreadsheets, which can further complicate access and interpretation of data as well as data discovery. For instance, the attribute labels of a data set may not be present in the data file, a single file or spreadsheet may contain multiple data sets, and integrity constraints are almost always missing. Metadata schemes such as the Ecological Metadata Language (EML) [5,18] provide standard ways of describing these implicit aspects of data sets.

In Figure 1, we show a fragment of EML for describing the 'wt' and 'LL' attributes of the data sets. EML is widely used within the ecological community for describing data, and provides support for explaining data set schemas (attributes, domains, and constraints) as well as the methods and protocols used to collect data, information about who collected the data and when, and access rights associated with data usage. While a large amount of (free-text) information is often stored within EML metadata, similar to other metadata standards the terms and application of terms within these descriptions are often unstructured and uncontrolled [5]. For instance, although EML can be used to represent the basic *structural* aspects of data—the number of attributes, their names, and their allowable values—the *semantics* of the data set—the types of entities observed, the characteris-

tics of these entities that were measured, and how these entities were observed in relation to each other—is either indirectly described (e.g., within the methods section of the metadata document) or are altogether missing.

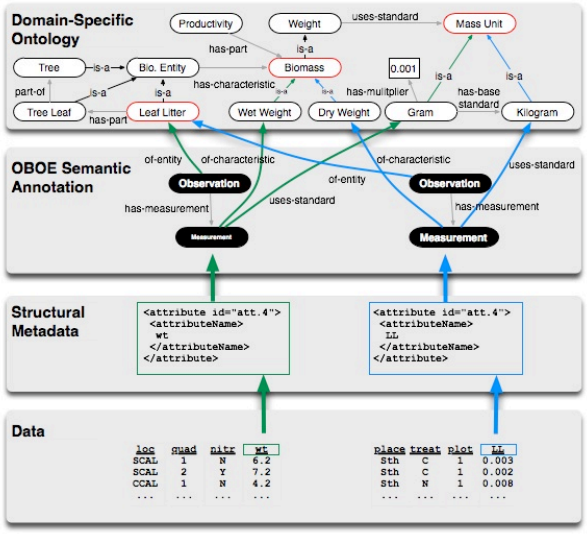

Figure 1. The components of our semantic-search framework including relational data, EML-based metadata, semantic annotations based on OBOE, and OBOE domain-ontology extensions.

The OBOE ontology [15,17] was developed to provide a rich set of concepts for describing the semantics of observational data. OBOE defines various OWL-DL classes and properties for representing and classifying observations and measurements. In OBOE, an observation consists of an observable entity (e.g., 'leaf litter'), a set of measurements, and a set of contexts that are represented through additional observations (e.g., a location or fertilization treatment). A measurement in OBOE consists of the characteristic being measured (e.g., 'weight') the measurement standard (e.g., a unit such as 'gram'), the measured value, and one or more qualifications such as precision. An OBOE extension typically represents a domain-specific ontology describing a limited set of concepts relevant to a specific organization, community, or group of researchers.

Semantic annotations extend EML by providing a mechanism to describe data set attributes in terms of OBOE concepts. An annotation is a formal structure, which represents a mapping from data set values to ontology instances (i.e., individuals) [17], and an XML-based syntax is used to represent annotation mappings. As shown in Figure 1, we can see that the two annotated attributes: (i) represent observations of leaf-litter entities; (ii) measure the weight of leaf-litter (although using different weight characteristics); and (iii) use compatible but different measurement units (kilograms and grams). We use annotations in a variety of ways. For instance, annotations can be used to find all data sets related to a particular concept, determine all of the concepts related to particular data set attributes, and compare data sets based on their corresponding OBOE structures (which can facilitate data integration).

A more detailed example showing the various XML syntaxes used for representing EML attributes (bottom), semantic annotations (middle), and an OBOE ontology extension (top) are shown in Figure 2.

Ontology

```
<owl:Class rdf:about="#RomaleaGuttatusPopulation">
  <rdfs:label>Romalea guttatus Population</rdfs:label>
  <owl:equivalentClass>
    <owl:Restriction>
      ....
    </owl:Restriction>
  </owl:equivalentClass>
  <rdfs:subClassOf rdf:resource="#GrasshopperPopulation"/>
</owl:Class>
```

Annotation

```
<sms:observation label="o4" distinct="yes">
  <sms:entity id="gce:GrasshopperPopulation"/>
  <sms:measurement label="m4a" precision="1" key="no">
    <sms:characteristic id="oboe:EntityName"/>
    <sms:standard id="gce:PopulationTypeNames"/>
    <sms:domain>
      <sms:entity id="gce:RomaleaGuttatusPopulation"/>
      <sms:entity id="gce:SchistocercaObscuraPopulation"/>
    </sms:domain>
  </sms:measurement>
</sms:observation>

<sms:map attribute="att.6" measurement="m4a"
        if="att.6 eq 'G6'" value="RomaleaGuttatus"/>
```

Metadata

```
<attribute id="att.6" scope="document">
 <attributeName>Species_Code</attributeName>
 <attributeDefinition>
    Coded species name
 </attributeDefinition>
 <storageType
   typeSystem="http://www.w3.org/2001/XMLSchema-datatypes">
   string
 </storageType>
</attribute>
```

Figure 2. Fragment of XML documents describing metadata, semantic annotations, and domain-extension ontologies. The annotation links the data attribute desribed in the metadata to the term defined in the ontology.

3. Semantic Search Strategies

As described in Section 1, our goal is to extend the KNB infrastructure to facilitate more effective data discovery by leveraging ontologies and semantic annotations. Here we briefly describe our extensions to the Metacat system [19] and the search strategies enabled by these extensions.

Metacat provides the metadata management services for the KNB as well as other EML-based metadata repositories employed by different institutions and research groups (e.g., South Africa National Parks, The Long Term Ecological Network, The Partnership for Interdisciplinary Study of Coastal Oceans, among others). Figure 3 shows the basic components of the Metacat architecture including the semantic extensions for supporting the dif-

ferent types of search described below [19].

Metacat provides services for storing and managing both XML-based metadata (e.g., EML documents) and the underlying data sets described via EML (e.g., by generating database schemas and loading data via the structural descriptions provided by EML). Metacat also provides additional services, e.g., authorization support (based on policies provided within EML documents), communication APIs (e.g., for loading and retrieving data and metadata), and basic keyword search of metadata.

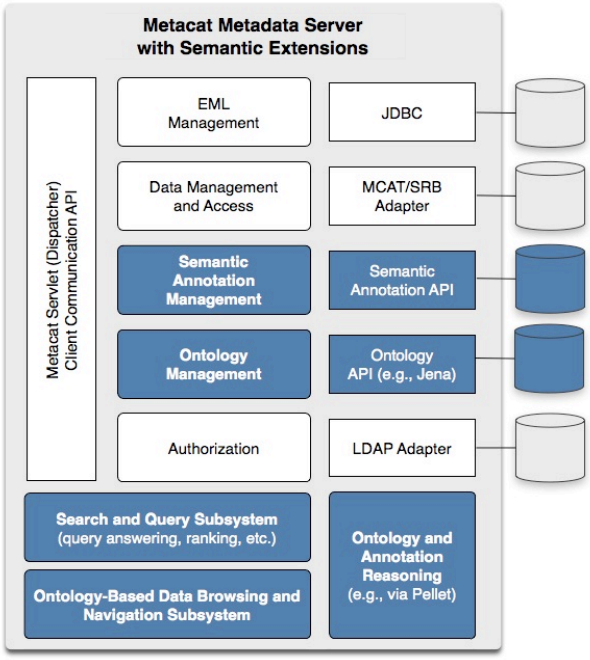

Figure 3. Metacat server architecture with semantic extensions (in blue).

As shown in Figure 3, we have added support to Metacat for storing and managing OWL-DL ontologies and semantic annotations, and for reasoning and search services to support different semantic-search strategies. Although not described here, we are also developing services within Metacat that use OBOE ontologies and semantic annotations to provide automated aggregate summaries of data (e.g., for browsing annotated data sets) and to support ontology-based data integration [15].

In our current implementation, the Jena API [20] is used to access ontologies and ontology terms within Metacat, and Pellet [21] is used to provide reasoning services over these ontologies (e.g., to compute class subsumption hierarchies and to ensure ontologies added to Metacat are consistent). We also extend Metcat's XML management capabilities with support for managing semantic annotations. Ontologies and annotations added to Metacat are assigned unique identifiers (URIs), allowing both to be easily accessed through external applications (e.g., Protégé [22]). Further, ontologies and annotations can be versioned using this URL scheme.

Metacat currently supports only XPath-based search that performs a string match of text contained within stored XML metadata elements. These keyword queries search over all text fields within metadata documents, returning results when terms match according to standard match rules given as part of the search, e.g., stating whether all keywords provided must match and whether exact keyword matches are required. This search can produce many false positives because the structure of the metadata is often ignored and only the text is searched. In addition (and as described above), many metadata documents contain information not only about the data that is being described, but also about the umbrella project or site that sponsored the data collection. Hence, a metadata document that describes a data set containing data on the distribution of zooplankton in lakes might have additional metadata describing the soil surrounding the lake. A search containing the keyword "soil" would return such a data set, even though the underlying data does not consist of any soil measurements.

The current implementation of Metacat's keyword search is also "agnostic" with respect to the relationships among terms, i.e., terms are treated purely as strings without any additional structure. For example, if two scientists identify the same specimen using two different (but essentially equivalent) species names, searching for one of these names will not result in a match for the other. A typical example within the KNB is when common or local names are used in place of a taxonomic name, e.g., resulting in searches for "*Romalea guttatus*" not returning data sets containing data described with the common name "grasshopper".

To improve overall precision and recall of Metacat searches we implemented several new search strategies.

Keyword-Based Term Expansion. In this approach, we "intercept" keyword queries issued to Metacat and expand them according to the term hierarchies of the stored ontologies. Specifically, if a given search keyword matches a class name (i.e., as specified by the *rdf:label* property of the class), then the search is expanded to include the synonyms of the class as well as the names of subclasses. This form of search alleviates the problem with simple keyword searches of not returning data sets described with synonyms or more specialized terms of the user-entered keywords. In our implementation, when a user enters a keyword search, Metacat locates synomyms and corresponding subclasses for each keyword using an ontology manager. The query is then augmented by Metacat with the expanded terms according to the given search constraints (i.e., whether terms should exactly match document terms and whether all given keywords must be present in a document) and executed against the current Metacat keyword search service. Although this strategy improves recall for documents that may have

been omitted with simple keyword searching, it can also cause additional false positives due to the addition of keywords. Thus, this approach generally increases recall, but not necessarily precision. Specifically, the set of metadata documents returned is always a superset of what is returned using the traditional Metacat keyword search.

Annotation-Enhanced Term Expansion. Semantic annotations allow individual data set attributes to be linked to one or more ontology classes. By applying keyword searches only to annotations, search results can potentially improve precision by returning fewer false positives. In annotation-enhanced search, each search term is first expanded using the ontology similar to the keyword-based term expansion. Here, when a search term matches an ontology class, we use the class and all subclasses to find matching annotations. An annotation is considered a match if it contains the corresponding classes according to the search constraints (see above). The metadata document linked to the matched annotation is returned by the search. For example, a metadata description of a data attribute described textually as "counts of grasshoppers" would be annotated as a "count per square meter" of "*Romalea guttatus*" by linking the attribute to the ontology classes that define these concepts. When the user searches for "grasshopper", the term is expanded to "*Romalea guttatus*" via the ontology's class hierarchy, and the annotation linking the metadata attribute to the class "*Romalea guttatus*" becomes a match.

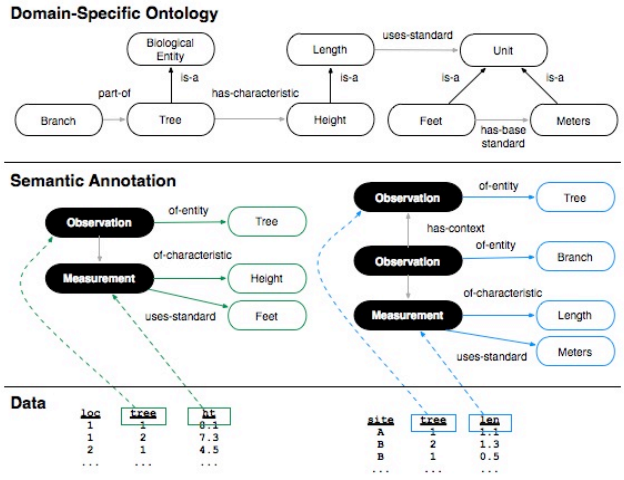

Figure 4. Example annotations demonstrating more precise search results for observation-based structured query.

Observation-Based Structured Query. Though the annotation-enhanced term expansion approach limits search to the relevant portions of metadata that describe data content (via attribute annotations), it does not take advantage of observation and measurement structures and relationships. In the observation-based query approach, users can search for data sets via their observed entities (organism, site, etc.) and the characteristics and standards used to measure them. Consider the example in Figure 4 in which two similar data sets are shown with their corresponding annotations. The first data set contains measurements of tree heights within different locations, whereas the second data set contains measurements of tree branch lengths at different sites. As shown in the example ontology of Figure 4, height is modeled as a subclass of length. Using the annotation-enhanced term expansion approach, searching for "tree length" would return both data sets. Although the second data set contains tree observations and length measurements, it would likely result in an irrelevant match for the given search.

Alternatively, in an observation-based search, queries are specified by explicitly filling in an observation "template" where ontology classes are given for the observed entity, measurement characteristic, and measurement standard (e.g., see Figure 5). To search for data sets containing tree lengths, we would fill in an observation query template, using the tree class as the observed entity type and the length class as the measurement characteristic type. Search is performed by finding matches between the observation types of annotations and the query template, where a match includes searching subclasses of the template classes. For the above example, the structured search would return only the first data set, thus providing a more precise search result.

Figure 5 shows a portion of the Web interface we have developed for accessing these Metacat semantic extensions. Different search approaches are selected using the search-type field. The interface also allows new ontologies, annotations, and EML metadata documents to be uploaded and viewed.

For observation-based queries, the interface is implemented so that only observation templates that return matches can be created . The Metacat extensions include support for computing the "active domain" of annotations, i.e., the ontology classes used by annotations. The interface shown in Figure 5 populates the set of choices for the observation-based query fields (entity, characteristic, and measurement standard type) according to the active domain. For example, if a user first selects an observed entity class, the choice of characteristic and standard types are narrowed to the corresponding sets of classes in the active domain that can result in a match. Thus, whenever one of these fields is changed, the list of choices for the other fields are dynamically updated, preventing users from choosing a combination that will not yield a result. This approach can further help users in more quickly discovering relevant data.

4. Discussion and Conclusion

We have described extensions to the Metacat system that provide new ontology-based services and approaches for

data discovery. Our goal is to develop generic, semantic-based approaches that will allow researchers to easily find and access relevant data. The approaches described here can improve precision and recall compared to the existing keyword-based queries offered by Metacat. As part of this work, we have also created domain-specific OBOE extensions for describing a small corpus of EML documents from the Georgia Coastal Ecosystems LTER site. As future work, we intend to broaden this initial set of ontologies and annotations to other EML documents stored within the KNB as well as perform formal precision and recall evaluations of our techniques. We also intend to extend our current implementation with more advanced structured searches (e.g., to search over observation context) and with capabilities for data browsing, summarization, and integration.

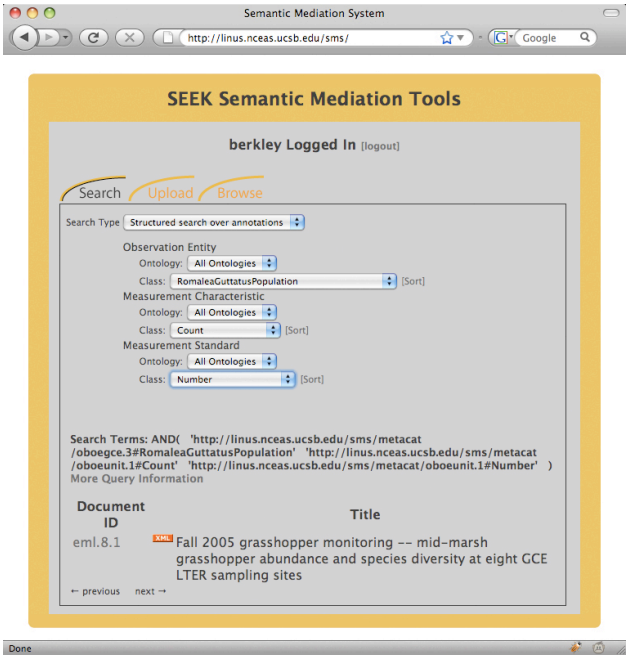

Figure 5. Web application for searching the new Metacat semantic-search extensions. Shown is an observation-based structured query and the search result.

5. Acknowledgements

This work is supported in part by the National Science Foundation under grants ITR 0225674, 0225676, and 0743429.

6. References

[1] R. Costanza et al., "The value of the world's ecosystem services and natural capital," *NATURE*, vol. 387, May. 1997, pp. 253-260.

[2] J. Green et al., "Complexity in Ecology and Conservation: Mathematical, Statistical, and Computational Challenges," *Bioscience*, vol. 55, 2005, pp. 501-510.

[3] B. Worm et al., "Impacts of biodiversity loss on ocean ecosystem services," *SCIENCE*, vol. 314, Nov. 2006, pp. 787-790.

[4] C.S. Parr and M.P. Cummings, "Data sharing in ecology and evolution," *TRENDS IN ECOLOGY & EVOLUTION*, vol. 20, Jul. 2005, pp. 362-363.

[5] M.B. Jones et al., "The new bioinformatics: Integrating ecological data from the gene to the biosphere," *ANNUAL REVIEW OF ECOLOGY EVOLUTION AND SYSTEMATICS*, vol. 37, 2006, pp. 519-544.

[6] W.B. Frakes and R. Baeza-Yates, *Information Retrieval: Data Structures & Algorithms*, Prentice Hall, 1992.

[7] B. Schatz, "Information Retrieval in Digital Libraries: Bringing Search to the Net.," *Science*, vol. 275, 1997, pp. 327-333.

[8] J. Rowley, "The controlled versus natural indexing languages debate revisited: a perspective on information retrieval practice and research.," *Journal of Information Science*, vol. 20, 1994, pp. 108-119.

[9] S. Brin and L. Page, "The anatomy of a large-scale hypertextual Web search engine," *Proceedings of the seventh international conference on World Wide Web*, vol. 7, 1998, pp. 107-117.

[10] N. Guarino, C. Masolo, and G. Vetere, "OntoSeek: content-based access to the Web," *Intelligent Systems and Their Applications, IEEE*, vol. 14, 1999, pp. 70-80.

[11] V. Uren et al., "The usability of semantic search tools: a review," *KNOWLEDGE ENGINEERING REVIEW*, vol. 22, Dec. 2007, pp. 361-377.

[12] D. Vallet, M. Fernández, and P. Castells, "An Ontology-Based Information Retrieval Model," *The Semantic Web: Research and Applications*, 2005, pp. 455-470; http://dx.doi.org/10.1007/11431053_31.

[13] T. Berners-Lee, J. Hendler, and O. Lassila, "The Semantic Web," *Scientific American*, vol. 284, 2001, pp. 34-43.

[14] J.S. Madin et al., "Advancing ecological research with ontologies," *TRENDS IN ECOLOGY & EVOLUTION*, vol. 23, Mar. 2008, pp. 159-168.

[15] J. Madin et al., "An ontology for describing and synthesizing ecological observation data," *ECOLOGICAL INFORMATICS*, vol. 2, Oct. 2007, pp. 279-296.

[16] M. Smith, C. Welty, and D. McGuinness, "OWL Web Ontology Guide.," 2004.

[17] S. Bowers, J. Madin, and M. Schildhauer, "A Conceptual Modeling Framework for Expressing Observa-tional Data Semantics," *Lecture Notes in Computer Science*, 2008.

[18] E. Fegraus et al., "Maximizing the Value of Ecological Data with Structured Metadata: An Introduction to Ecological Metadata Language (EML) and Principles for Metadata Creation.," *Bulletin of the Ecological Society of America* , vol. 86, 2005, pp. 158-168.

[19] M.B. Jones et al., "Managing scientific metadata," *IEEE INTERNET COMPUTING*, vol. 5, Oct. 2001, pp. 59-68.

[20] "Jena: A semantic web framework for Java."; http://jena.sourceforge.net.

[21] E. Sirin et al., "Pellet: A practical OWL-DL reasoner," *Journal of Web Semantics*, vol. 5, Jun. 2007, pp. 51-53.

[22] H. Knublauch, M.A. Musen, and A.L. Rector, "Editing description logic ontologies with the Protégé OWL plugin.," *CEUR*, vol. 104, 2004.