Ontologies for ecoinformatics

Richard J. Williams [a,b], Neo D. Martinez [a], Jennifer Golbeck [c,*]

[a] *Pacific Ecoinformatics and Computational Ecology Lab, 1604 McGee Ave. Berkeley, CA 94703, United States*
[b] *San Francisco State University, Computer Science Department, San Francisco, CA, United States*
[c] *University of Maryland, Computer Science Department, A.V. Williams Building, College Park, MD 20742, United States*

Abstract

Rapid advances in information technologies continue to drive a flood of data and analysis techniques in ecological and environmental sciences. Using these resources more effectively and taking advantage of associated cross-disciplinary research opportunities poses a major challenge to both scientists and information technologists. These challenges are now being addressed in projects that apply knowledge representation and Semantic Web technologies to problems in discovering and integrating ecological data and data analysis techniques. In this paper, we present an overview of the major ontological components of our project, SEEK ("Science Environment for Ecological Knowledge"). We describe the concepts and models that are represented in each, and present a discussion of potential applications of these ontologies on the Semantic Web.

Keywords: Ecoinformatics; Ecology; Ontologies; Semantic Web

1. Introduction

Rapid advances in information technologies continue to drive a flood of data and analysis techniques in ecological and environmental sciences. Using these resources more effectively and taking advantage of associated cross-disciplinary research opportunities poses a major challenge to both scientists and information technologists. For example, unlike DNA sequences collected by molecular biologists, raw data collected by environmental biologists are rarely made available to many other scientists. If the data are made available, access typically requires days if not weeks of human labor and years of time before other scientists can analyze the data. More often, interested scientists are unaware that such data even exist and the data are more or less lost to history. These challenges have been recently addressed by projects such as "Science Environment for Ecological Knowledge" (SEEK) and "Semantic Prototypes in Research Ecoinformatics" (SPiRE), which apply knowledge representation and Semantic Web technologies to problems in discovering and integrating ecological data and data analysis techniques. These technologies rely on ontologies that appropriately capture and encode scientific knowledge from the domains of interest.

The SEEK and SPiRE projects have developed a collection of ontologies for describing ecological organisms, systems, and observations. The two major uses of ontologies are accessing and analyzing ecologically important information. The first activity uses the ontologies to describe ecological and environmental data sets in sufficient detail to permit automation of the discovery of data sets relevant to addressing a particular scientific question. The second activity uses the ontologies to describe data analysis tools so that the semantic mediation system can assist in the selection of tools and creation of scientific workflows given semantic descriptions of the incoming data and/or the desired results.

The ontologies described here were designed to provide a rich description of ecological and environmental data sets, so that the first of these uses could be accomplished. Relevant characteristics of a data set that need to be described using ontologies include (1) where, when, and by who the data were collected, (2) a description of what was observed, typically including the taxonomic classification and other traits of observed organisms, and (3) the sampling protocol, including collection procedures and associated experimental manipulations.

The ontologies are written using OWL and are contained in a number of separate OWL documents. Conceptually separate parts are contained in individual files. Together, the files describe

* Corresponding author. Tel.: +1 301 314 6604; fax: +1 301 314 9734.
E-mail addresses: rich@sfsu.edu (R.J. Williams), neo@PEaCElab.org (N.D. Martinez), golbeck@cs.umd.edu (J. Golbeck).

a broad model of information covering various domains of interest, which is then used to develop more highly domain-specific models. The ontology currently addresses two broad areas, scientific observations and ecological and environmental science. Several sub-models describe scientific observations and data sets including models of space, time, units, and dimension as used in describing data sets. The ontologies were developed using the OWL-DL variant of the OWL languages, and all the ontologies are available online at http://wow.sfsu.edu/ontology/rich/.

This paper will present overviews of some of the major ontologies developed for the SEEK and SPiRE projects, and describe the concepts and models that are represented in each. We conclude with a discussion of potential applications of these ontologies on the Semantic Web.

2. Ecology ontology

A diagram of a portion of the EcologicalConcepts.owl class hierarchy is shown in Fig. 1. While omitting much detail, this figure serves as a useful orientation to the concepts discussed in this section. Ecology is modeled as the study of a system's components and the changes over time in those components. The

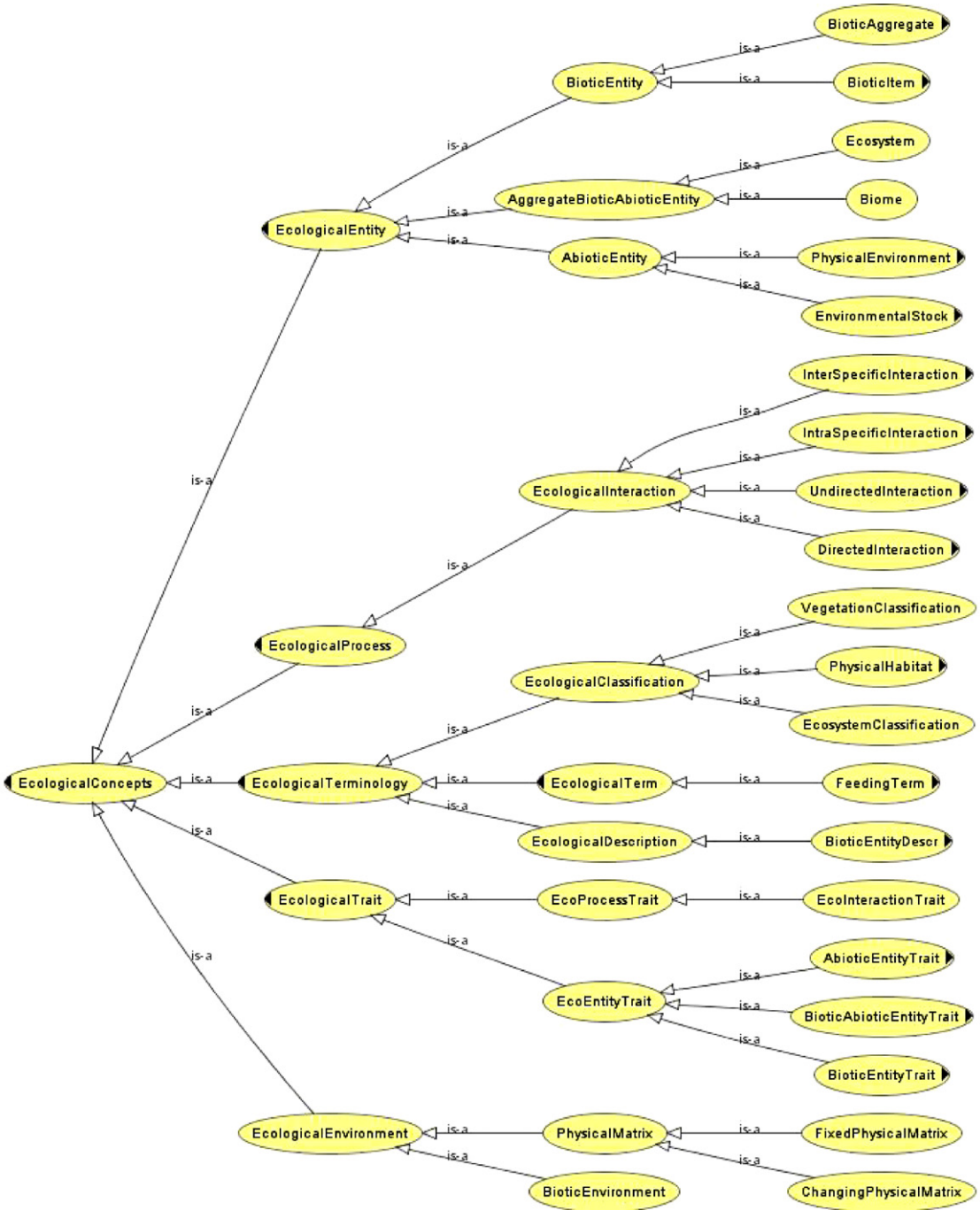

Fig. 1. A portion of the EcologicalConcepts.owl class hierarchy.

system components are typically biotic entities, though sometime abiotic entities are also seen as part of the system of interest. Changes in the system occur due to processes that affect the entities in the system, interactions between the entities, either biotic or abiotic, that make up the system, and interactions between the system's components and its environment.

Biotic entities are separated into two basic types, either individual organisms or some kind of group or aggregate of organisms. Individuals and groups are in turn are divided into whole organism (s) or parts of organisms. Groups are often aggregated at the level of species, but ecologists also use many other types of aggregation. Among these are taxonomic groupings resolved to something other than the species level (e.g., kingdom, phylum, order) including organisms referred to by common names (e.g., dogs, plants) rather than taxonomic designations, organisms with the same functional role in the ecosystem (e.g., carnivores), and organism parts (e.g., leaves).

Interactions between entities are modeled either as directed, such as a predator–prey interaction, or undirected, such as competition. Directed interactions are further subdivided into interactions in which there is material exchange between the interacting entities, such as the energy and nutrient exchange that occurs during a predator-prey interaction, and interactions in which entities exchange information, such as chemical or audible signals.

The traits of entities and interactions that a scientist chooses to observe are typically influenced by the scientist's theories and hypotheses. The same is true of the way in which the scientist subdivides the observed world into individual entities and interactions between those entities. However, scientifically interesting traits that capture scientists' interest typically change more frequently than the entities possessing the traits. The ontology is designed around this understanding, defining properties independently of entities and processes. This flexible architecture allows new properties to be defined and attached to entities and interactions without having to refine the underlying entity and interaction classes.

An ecological system of interest typically comprises one or more entities that exist within an environment. An Ecological Environment describes the environment of one or more entities. The environment is itself one or more biotic or abiotic entities, as the distinction between entity and environment is dependent on the perspective of the researcher. The distinction is that in a particular experimental context, the environment is seen as external to and not influenced by the system, but the environment might influence the system.

Descriptive terms are frequently attached to ecological entities or interactions. For example, a scientist might label a species as a predator or an omnivore. Descriptive terms, such as terms used to describe the feeding behavior of various organisms, are difficult to model and categorize because they often do not fit cleanly into a set of independent categories. A second problem is that the terms are sometimes ambiguous and there is not clear agreement among domain scientists to a term's meaning. Despite these problems, descriptive terms are widely used and can give important insight into a scientist's understanding of the ecological role of the entity in question.

By analyzing the terms used in the ecological literature to describe feeding behaviors, we identified a small set of independent underlying concepts that can be used to define many of these terms. For example, consider the terms predator, parasite and omnivore. Omnivore is inconsistently defined in the scientific literature. Ecology texts define omnivores as organisms that consume species that are at different trophic levels in the food web (for example, see Ref. [1]). However, most biology texts define omnivores as organisms that regularly consume both plant and animal taxa (for example, see Ref. [3], http://www.biology-online.org/dictionary/omnivore). We address this by separating omnivores into trophic omnivores and its subset taxonomic omnivores. The different definitions of omnivore mean that a term-based query, such as "find all the omnivores in a food web" will have different results depending on the definition of omnivore used, and that the definition of omnivore expected be an ecology researcher and a high school biology student are different. Using terms whose definitions are precisely specified within the ontology will help resolve these ambiguities.

Another useful example is the distinction between the terms predator and parasite. "Predator" refers to an animal that kills and eats other animals. "Parasite" also refers to an animal that consumes some species, termed the host, but this interaction occurs for an extended period of time and usually does not kill the host. Thus, predators and parasites can be distinguished by both the relative duration of the interaction between the two organisms and whether that interaction results in the immediate death of the organism being consumed.

These and many other idiosyncratic descriptions of feeding behaviors are systematically broken into several independent components concerning what is being eaten, how it is eaten, and the effect of the eating. These components include taxonomic categorization, trophic level, the part of the prey consumed, the relative duration of the feeding interaction, and whether the feeding interaction leads to the death of the prey. The meaning of feeding terms can be captured by one or more of these descriptors and terms may have more meaning than is currently captured in our feeding ontology. Still, the ontology provides a useful and relatively rich way of defining many of the terms used to describe feeding behaviors.

3. Ecological models, analysis methods, and ecological networks ontologies

Ecology is modeled as a science of entities and interactions occurring within an environment. This structure greatly facilitates the description of the many models in ecology. The model ontology allows the rules of the interactions of entities to be specified. Models are categorized using the common ecological divisions of individual, population, community and ecosystem, with further divisions within each of these broad categories.

Models are considered to be composed of model entities, interactions and parameters. Each of these is further subdivided like the underlying model, into individual, population, community and ecosystem model concepts. Parameters can be associated with entities or interactions.

A variety of interaction networks are studied by ecologists, the most familiar being food webs, which records who eats whom in an ecological community. Networks connect nodes with either directed or undirected links. A food web is a network with directed links (DirectedNetwork), where each node represents a set of organisms and each link represents a feeding interaction, in which biomass and therefore energy is transferred from some prey or otherwise classified organisms to the consumer organisms.

In addition to a set of entities associated with food webs, there is also a set of properties associated with the entities. Properties associated with a Food Web as a whole include connectance, the number of links and the number of species. Properties associated with a node in the food web include its trophic level, connectivity, generality and vulnerability.

4. Observations and measurements ontology

Concepts used to describe scientific observations and measurements, independent of a particular scientific domain, are defined in MeasurementBase.owl. Our model of the empirical scientific process of measurement is based on two fundamental concepts: the observation and what is being observed. The observation can be of an item or its trait; in other words, a scientist either observes the existence of something or measures some trait of the entity being observed.

A trait links two instances, the Trait that was observed and the entity that the Trait was measured on. By decoupling the entities and traits, it is possible to extend the ontology by introducing a new trait without having to change the definition of the entity that was observed. This accommodates the typical mode of innovation in scientific research, in which novel traits are commonly developed, whereas the entities that the scientist studies, while also evolving as scientific understanding develops, change much more slowly. The value of a trait frequently has units associated with it, and so there is a separate ontology describing units and dimensions.

Observations are typically made at a particular time and location, and using a specified measurement procedure. We refer to these as the context of the observation. The ontology includes classes and properties to express knowledge about the temporal and spatial context of a measurement. Descriptions of spatiotemporal regions can include potentially nested intervals in both space and time. The ontology also allows simple descriptions of measurement procedures. Understanding whether data sets can be integrated depends on large part on understanding whether measurement procedures are compatible. Automating data integration will require a detailed ontology of ecological measurement procedures.

5. Additional ontologies

Hierarchical classification and taxonomic identifiers are important for organization and identification. SEEK includes a simple ontology for representing taxonomic identifiers that can have the seven main taxonomic ranks[1]. Instances of ranks point to higher and lower ranks so that the hierarchy of ranks can be traversed. The names of taxonomic ranks can be stored locally or can be referenced to an external data file.

Ecological niche models are a category of ecological data analysis methods that are used to model the spatial distribution of species. The input data are the locations of the existence of an organism, and environmental conditions at those locations. The output is a prediction of the niche of the organism, or the conditions under which the organism can survive. To support this modeling, the EcologicalNicheModeling ontology contains extensions of ecological models to describe the analysis techniques and the concept of an ecological niche.

6. Problems and limitations

As mentioned in the introduction, we chose to use the OWL-DL sublanguage. Using OWL-Full was considered, as there were several places where using classes as instances might have been a useful modeling construct. We chose to avoid this construct and find other ways to express the concepts that might have used this construct so as to stay within the confines of OWL-DL and be able to take advantage of the reasoning tools this language offers.

There are many places in the ontologies where additional relationships between classes and constraints on instance values are known but cannot easily be expressed using OWL-DL. Many class definitions could be made more precise by adding more complex property value restrictions than are easily expressible in OWL, such as restricting the class of an instance that is the value of an object property that itself is a property of an object property of the class in question, or restrictions on similar but longer chains of property values. Another potentially useful property value restriction not expressible in OWL is a restriction on the range of values of a numeric datatype property, such as requiring that the property be less than some value. A simple example of numerical reasoning is that a geospatial location instance could be inferred to be of class TropicalLocation if the value of its latitude property was within 23° of the equator. Classification based on numerical value of properties is common in science and is an important omission in the current OWL languages. These relationships will be added to the ontologies using a rule system such as the proposed Semantic Web Rule Language.

Another potentially useful property value restriction difficult to express in OWL are restrictions on the range of values of a numeric datatype property, such as requiring that the property be less than some value. A simple example of numerical reasoning is that a geospatial location instance could be inferred to be of class TropicalLocation if the value of its latitude property was within 23° of the equator. While user-defined XML Schema datatypes can be used to define such a range and referred to from an OWL document by a URI, there is no standard mechanism for generating a URI for a particular user-defined datatype[2] [ref:

[1] Kingdom, Phylum, Class, Order, Family, Genus, and Species.

[2] http://www.w3.org/TR/swbp-xsch-datatypes/.

http://www.w3.org/TR/swbp-xsch-datatypes/]. Thus, the ability, while technically there, is neither portable nor widely supported by OWL tools. Classification based on complex relations between numerical values of properties is common in science, and we anticipate needing relations more complex that even offered in the OWL 1.1[3] proposal. These relationships will be added to the ontologies using a rule system such as the proposed Semantic Web Rule Language.

There are several places in the ontologies where classes are related mathematically. A simple example occurs in the specification of the dimension of a measured value. The dimension acceleration is specified as an instance of the DerivedDimension class and is composed of two DimensionPart instances that are multiplied together, distance and time^{-2}. While the DerivedDimension class has a dimensionParts property that must have one or more DimensionPart instances as values, the ontology does not include the information that these DimensionPart instances must be multiplied together to form the dimension. The existence of unspecified relationships between various classes means that the OWL specification is incomplete and any user of the ontology must have knowledge of these unspecified relationships in order to make full use of the ontology. This information is currently included as comments in the relevant classes.

7. Applications

The ontologies described here are designed to facilitate the accessing and sharing of information about ecological systems on the Semantic Web. The ontologies are used in a food web knowledge base being developed as part of the Webs on the Web project [5]. Food webs are models of trophic relationships in an ecosystem. They are built up from observations about what species are found in an ecosystem and what those species eat there. This data is compiled from studies of individual species, including direct observation of feeding interactions and examination of stomach contents. However, it is difficult to directly observe all components of a species' diet. As a result, the effort required to assemble a food web is very large and many food webs have been criticized for being incomplete or inaccurate. The individual nodes in a food web, here called species, are in fact often groups of functionally similar species, and sometimes are only identified with their common name rather than any taxonomic specifier. More complete and accurate food webs would be valuable both for fundamental scientific research into the dynamics of complex ecosystems and for application in the conservation and management of ecosystems.

Various applications of the highly inter-related data represented in the knowledge base are under development. Some data sets contain brief descriptions of the feeding habits of the various species. For example, the species might be tagged as being an herbivore. Using the definition contained in the ontology that an herbivore is an organism that consumes plants, this information can be used to test the consistency of the food web data, namely that all items consumed by the herbivorous species must be members of the plant kingdom. If the same species appears in a different food web, the fact that it is known to be an herbivore can still be used to check the consistency of the feeding links in the new web. If in inconsistency is found, there is either an error in the food web data or an error in the original determination that the species in question is an herbivore. The categorization that a species in a food web is an herbivore can also be used to infer missing taxonomic information about that species' prey when only the prey's common name is given (that common name can be inferred to belong to a plant) or to perform concept-based queries, such as locating all herbivores in a particular food web.

Another application under development is an effort to infer trophic relationships between species in a food web when observational data beyond the co-existence of species in a habitat is lacking. We have named this project Meal of a Meal (MOAM) in a nod to the Friend of a Friend (FOAF) project that is used in the Semantic Web representation of social network data [2]. Using data about food webs available on the Semantic Web, as well as biological taxonomies and phylogenies, also represented with Semantic Web ontologies, the goal of MOAM is to integrate this information into a single model, and develop algorithms that will suggest possible trophic relationships. The methods we are developing use taxonomic and phylogenic similarity measures between species, known diet similarities of related species, data on relative body sizes and data on habitat similarities to infer trophic connections. The logic supporting this is (1) that many if not most species who are closely related phylogenetically eat similar diets, (2) the diet similarity of taxonomically related species varies across taxa, (3) species diets are strongly constrained by the relative body size of predator and prey [4], and (4) inference is more likely to be accurate when closely related species used to infer diets live in similar habitats to the target species.

8. Conclusion

We have described a collection of ontologies for representing scientific ecological data on the Semantic Web. The ontologies are not only models of the connections between ecological concepts, but are also used in several applications to represent instance data. Ecology, as a field where extracting useful data from others' research is a challenge, is a prime example of were Semantic Web-based data sharing is potentially valuable. Currently, this work is largely independent of the web, other than as a distribution mechanism. The food web instance data used in the WoW project is collected into a single centralized knowledge base rather than being drawn from a distributed set of data. This is mainly a result of the fact that the semantic web and associated tools are in their infancy and the ecological community is not currently producing data with semantic markup and publishing it on the semantic web. We see this work as a step on the path to that goal. These ontologies and the applications that use them serve as a case study of how carefully crafted scientific ontologies can be used

[3] http://www-db.research.bell-labs.com/user/pfps/owl/overview.html#2.3.

to facilitate application development and data sharing on the web.

Acknowledgements

This work was supported by supported by NSF projects ITR-0326460 (SPiRE) and ITR-0225676 (SEEK) as well as Fujitsu Laboratories of America—College Park, Lockheed Martin Advanced Technology Laboratory, NTT Corp., Kevric Corp., SAIC, National Science Foundation, National Geospatial-Intelligence Agency, Northrop Grumman Electronic Systems, DARPA, US Army Research Laboratory, NIST, and other DoD sources.

References

[1] M. Begon, J.L. Harper, C.R. Townsend, Ecology: Individuals, Populations and Communities, third ed., Blackwell Science, Oxford, 1996, p. 1068.

[2] D. Brickley, L. Miller, "FOAF Vocabulary Specification", July 27, 2005 <http://xmlns.com/foaf/0.1/>.

[3] N.A. Campbell, J.B. Reece, L.G. Mitchell, Biology, fifth ed., Benjamin/Cummings, Menlo Park, CA, 1999.

[4] J.E. Cohen, S.L. Pimm, P. Yodzis, J. Saldana, Body sizes of animal predators and animal prey in food webs, J. Anim. Ecol. 62 (1993) 67–78.

[5] I. Yoon, R.J. Williams, S. Yoon, J.A. Dunne, N.D. Martinez, Interactive 3D visualization of highly connected ecological networks on the WWW. ACM Symposium on Applied Computing (SAC 2005), Multimedia and Visualization Section, 2005. pp. 1207–1217.