EnviroInfo 2005 (Brno)
Informatics for Environmental Protection - Networking Environmental Information

Overcoming the Multiple Islands of Ontologies

Schleidt Katharina, Schentz Herbert, Mirtl Michael

Abstract

One of the major challenges facing ecologists today is the problem of obtaining sufficient amounts of information pertaining to widespread ecological phenomena. While the physical information and data networking does not pose a great challenge using today's technologies, the difficulties start when one tries to describe the metadata pertaining to the data in such a manner that other users of this data can exactly ascertain how the data was collected (where, how, who) and how it may be validly analyzed. In order to structure the available data and metadata, one is turning to the use of ontologies.

As ecology, like all sciences, is not a static field, new ideas, methods and procedures are continually being introduced. This means that an ontology designed to structure data for this field must be extendable in order to adapt to emerging requirements.

In order to meet these needs, we propose a procedure for the collaborative creation and extension of ontologies. By further splitting the ontology into a (relatively static) core ontology supplemented by various domain ontologies, it is possible for multiple communities to independently extend their domain ontologies without disrupting work on other domain ontologies. Thus, a flexible but stable system can be defined which meets the needs for dynamic extension while retaining the stability required for a large scale network.

1. Ontologies for Ecological Data Networking

As ecology is a very wide field, a great many different types of data must be made available to allow for the discovery of the various relations between the multiple factors involved in ecological phenomena and for the development of models. Scientists and environment controlling institutes always need data of many others in addition to their own to assess the state of the environment.

Whilst to date several attempts have been made to create such networks for ecological data, most of these have been based on XML, defined by XML schema. This has been necessary for lack of alternatives, XML and XML schema are well introduced technologies, many tools and libraries are available. However, this is a quite inflexible form of data structuring, as once a schema has been defined it can only be changed or extended by introducing a new version of the schema, requiring a great deal of installation and configuration work for everyone who uses that schema. Using object oriented datamodels like i.e. OWL, the communication schema is not changed for new or changed information: classes are added or changed and broadcasted. Whoever needs the extended information uses it, whoever does not need it, needn't be bothered.

One of the major advantages of ontologies for the structuring of shared data is the feature of inheritance. This means, that when a new, more complex data type is required, this type can be derived from an existing datatype, retaining all existing fields while adding those additional fields required. This new datatype, while containing addition information, can still be used by systems expecting the old datatype, admittedly with data loss, but allowing access to the information that was contained in the parent type. This allows for greater flexibility in the extension of the data structures required.

Several years of experience in implementing such domain ontologies by use of MORIS (ev. reference...) have lead to a basic set of conceptual elements and strategies, tackling *inter alia* the challenges of representing and linking reference lists (taxonomy), application of a backbone ecosystem model throughout the system and across disciplines as well as developing a practicable set of object classes.

2. Necessity of Standards

The current status in this field is that each community is defining its own ontology without considerations of compatibility with those used in other communities. **It seems to be much easier to create an ontology than to read and understand it.** Although this allows for data exchange and sharing within special communities, it does not provide the means to do so on a general level and forces organizations to map their data to several different ontologies according to the communities in which they are taking part. While it is possible to define relations between certain components of one existing ontology to a separately defined ontology, one usually comes across parts that are extremely difficult or impossible to map. This happens, for example, when information that is defined as required within one ontology is either optional or non-existent within the other or when m:n relations exist between the frames of reference of one ontology and those of the others. These contradictory community ontologies often force organizations and scientists to decide on which community they want to share data with and to organize their data in such a way that it can be mapped to one specific ontology. This is contrary to the requirements of ecologists, as they need to be able to share data with as many others as possible, which in turn requires interoperability with many different communities. In order to facilitate this interoperability, it is necessary to define standards for ontologies. These standards should, where possible, be based on existing scientific, semantic, data transfer and other formal IT-standards.

Despite the existence of formal representation languages and many initiatives geared at standardization, we are still far from having a standard ontology for the representation of ecological knowledge as well as the sharing of data and applications

3. Need for an Extensible Ontology

The challenge being faced today is the definition of a widely accepted standardized ontology. Whilst the creation of a network of basic concepts is relatively simple, getting the various involved parties to commit to this schema will be a bit more difficult but possible. The real difficulty faced is the creation of a core ontology with high and widespread commitment, that can then be successively extended to nearly all the specialized needs of ecologists (domain ontologies).

3.1 Core Ontologies

To avoid chaos in the construction of ontologies, it is very important to have a stable "nucleus", a so called core ontology. A core ontology is that part of an ontology that represents the most important basic concepts and relations used by a community. All partners must commit to the core ontology and there must be agreed rules for its extension. This standardized core ontology must be designed in such a manner as to allow for flexible extension into various specialized domains.

A very important characteristic of ontologies is that they consist not only of terms but also of relations between terms; thus, a core ontology includes basic relations. There should not be too many relations in a core ontology.

We recommend including the following basic relations within the core ontology:

Relation	Description
IS A	basic relation of object orientated inheritance
HAS A	basic relation of attributes

PERTAINS TO	basic relation of object orientated membership to a class
IS PART OF	
HAS AS PARTS	inverse relation of IS PART OF
LIES ON (IN)	locality determining relation

3.2 Domain Ontologies

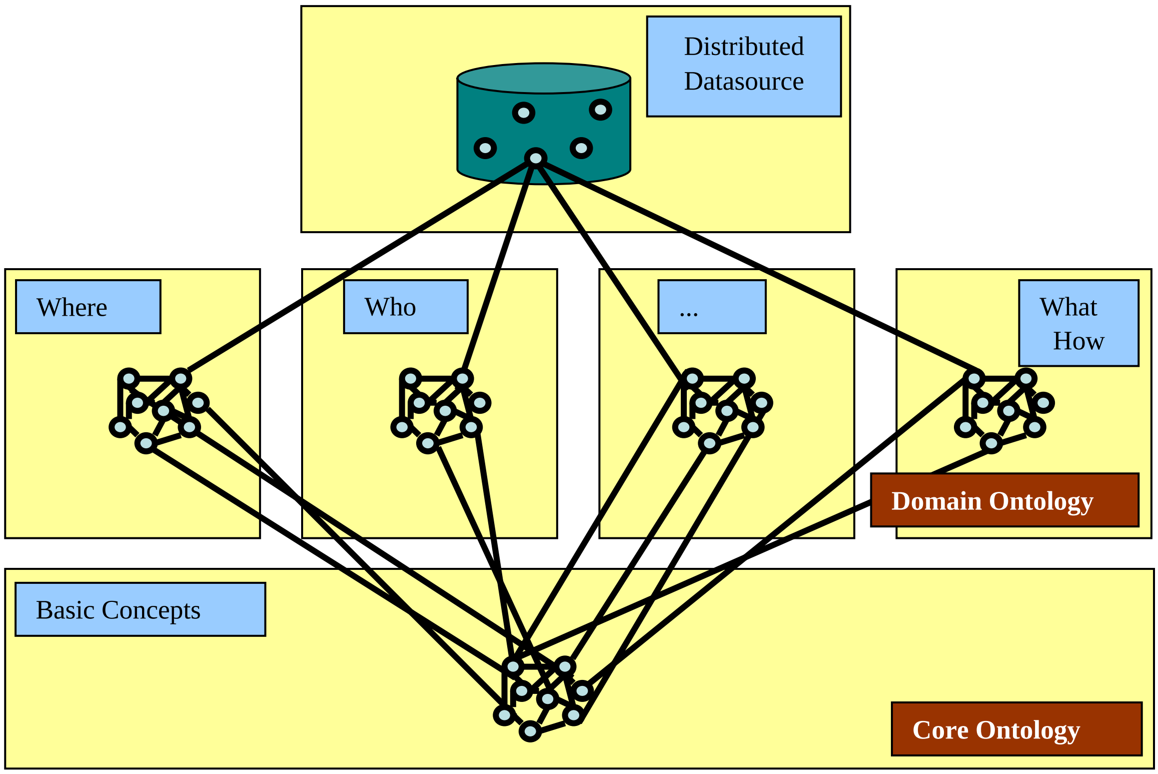
A domain ontology consists of that part of the concepts and relations that are necessary for the description of certain domain such as "terrestrial ecology", "biodiversity", "aquatic ecology" amongst the whole ontology for ecology.

A domain ontology is derived from a core ontology which delivers the basic concepts that are required by all or most domains. This helps to assure interoperability between individual domain ontologies based on the same core ontology.

The advantage of this division into a central core ontology as a basis for all further developments and domain ontologies for topic specific concepts is that the domain ontologies can be extended virtually independently of the other domain ontologies whilst retaining compatibility to the rest of the ontology family.

3.3 Distributed Ontologies

It does not make sense to develop one large ontology for the whole of ecology, or even worse for whole science in one giant step. As usual with software development and development of thesauri, it is strongly recommended to construct the ontologies out of several parts.



The maintenance of instances may be performed by different partners, i.e.:

TOPIC (with several classes)	**INSTANCE maintenance**
Taxa	GBIF, TDWG
Political units (states, districts, …)	Special institutes in every country
Geological regions	Geological Institutes
River basins Ground water bodies	Hydrographical Institutes of every country
Site descriptions	Site owner
Sample descriptions	Site owner
Observation results	Site owner

Handling of dislocated, distributed data:

- **Central maintenance, central availability:** The whole community is responsible for the maintenance and change of this part of the ontology. It is always available for querying and semantic mediation work. All classes and some instances (e.g. units) are represented in this part.
- **Dislocated maintenance, central availability:** Those are the parts maintained by certain specialists but must always be avail-able for querying and semantic mediation work.
- **Dislocated maintenance, available on demand:** Those are the parts that are maintained locally and will be harvested by the search engine on demand. It corresponds with the parts where the authority almost completely lies with the site owner.

4. Collaborative Creation of an Ontology Family

A first approach in the definition of basic ontologies is the attempt to merge existing thesauri, concept schemas and ontologies. The sole application of this method has rarely proved successful. This has lead us to the conclusion that a better way of effectively creating ontologies with a widespread commitment as well as coverage of the necessary range of topics is through the collaborative creation of a core ontology, incorporating existing top level ontologies as contributions. Even more important is the definition of a process for the subsequent collaborative extension of this core ontology to specific domain ontologies.
The following steps must be included in a defined procedure for the creation of an agreed upon core ontology and the collaborative extension to individual domain ontologies:

- Establish a voting mechanism
- Establish a search mechanism for concepts and relationships
- Establish a core ontology
- Define extension rules for the domain specific extensions including a versioning system
- Establish a system for the introduction of new concepts and relations

4.1 Voting Mechanism

In order to collaboratively create and extend an ontology, a voting mechanism, allowing users to show concurrence to proposed concepts and relations, must be established. Registered Persons, Organizations and Communities should be able to commit to concepts and relationships.

As this decision and coordination process make take quite a bit of time before new concepts are finalized, it will be necessary that concepts and relations that have not been finalized be properly flagged to show their status. This not only supports the voting mechanism itself, it allows users to utilize new concepts and relations before they have been finalized, based on their own judgment of the status of the voting process.

4.2 Search Mechanism

A search mechanism must be established allowing users to browse through existing and proposed concepts and relations. This search mechanism should allow the user search not only by the actual contents of the concepts and their relations but also by their acceptance status as well as by who has already voted what way on the object. (extended reasoning)

4.3 Establish Core Ontology

As a solid and stable core ontology is the necessary basis for all further domain ontologies, this step must be performed first. In order to do this work effectively, it makes sense to base this on the merge and mapping of various existing top level ontologies, thesauri and other contributions. This process will take a lot of discussion and mediation, but is very important as a high level of commitment must be attained between all parties involved.

4.4 Define Extension Rules

Once the core ontology has been created and finalized, a decision must be reached as to which topics require their own domain ontologies. Certain factors must be taken into account during this process:

- ♦ If the topic is too large, maintenance can become quite difficult as too many sometimes conflicting needs must be covered; the agreement of many parties must be attained.
- ♦ If the topic area is too small, the danger of overlaps with other domain ontologies becomes greater. This in turn leads to the danger of multiple definitions of the same concept, as the people working in one domain often will have little contact with those working on other domains.
- ♦ Domains should be precisely defined and the overlaps with other domains highlighted. Where there is interaction and cross-referencing between the different domain ontologies (i.e. references to persons, citations, …) the stand alone domain work ends and coordination and voting starts.

A decision must be reached as to what parts of the ontology may be easily extended, and which parts may only be extended after a lengthy agreement process. One suggestion of this is to allow the creation of decentralized instances, while requiring agreement for the definition of new classes.

4.5 Introduction of new concepts

Finally, a system for nomination of new concepts and relations for the domain ontologies must be established. This should include information such as who may recommend what sort of extension to which domain ontology.

5. Conclusion

In order to effectively share ecological data in a network environment, the use of ontologies can be quite effective. In order to attain widespread commitment to such an ontology, it must be created and extended collaboratively.