Morten Hulden

CWR Ontology

Project Report

A Practical Approach on Creating a Restricted Ontology for Crop Wild Relatives

2007

CWR Ontology – Project Report
A Practical Approach on Creating a Restricted Ontology for Crop Wild Relatives

Morten Hulden

May 30, 2007

Abstract

The task to identify a subset of about 400 terms highly relevant to crop wild relatives was performed as a continuation to an earlier project where a set of about 11400 term was extracted from on-line sources.

Terms with high relevance were grouped into themes, roughly corresponding to Agrovoc (top level) categories, or indicatives of the thematic sources from which the terms were collected (biological, geographical on-line dictionaries etc), with an attempt to balance the number of terms between the groups. For the import into the ontology structure the themes were converted to namespaces in order to preserve the grouping and allow manipulation within ontology client programs on terms based on namespace grouping. Before the import the namespaces were slightly modified and adapted to some other existing ontologies.

In addition to selecting relevant terms and definitions, definition of vertical and horizontal relationships between the terms was performed. Terms were also linked to sources (URIS) through Dublin Core extensions of the ontology structure.

The export from SQL to RDF/OWL was done with a script written in Perl that extracted the terms, descriptions, sources from the database, plus vertical hierarchy, term synonyms and other variants, as well as some simple horizontal relationships and produced an import file for Protégé in RDF/OWL format.

The resulting subset of terms was provided as a number of files; a main file containing core structure, object and data type definitions and term data in separate files per namespace, suitable for import by an ontology client program such as Protégé.

Introduction

The CWR ontology project is part of the CWR project, which involves both FAO and Bioversity International and other national stake holders.

The starting point for the CWR ontology was a larger set on 11407 terms that were extracted from on-line sources (glossaries, dictionaries, PDF-publications and thematic web-pages), during an earlier project. The terms have been stored in a relational SQL database.

Term definitions

The results from the previous project on CWR-related terms was used as the starting point. All information from the previous project, including copies of the original URI content had been stored in the SQL database. From these results (11407 terms) the fields containing terms, definitions, URIS and dates of last access were extracted from records flagged as glossaries or dictionaries, and were collected into a new database table, containing about 3500 terms. Some additional terms and descriptions were later added to this table manually later.

Terms

The target number of terms relevant to CWR was set to about 400 in this project. From the terms for which definitions existed in the database a very small set of terms was manually selected, roughly consisting of balanced subsets of terms representing the themes agriculture, botany, environment and protection, earth and soil sciences, genetics, law and resource management. In the database created during the previous project fields flagging the terms according to theme were used for creating the initial balanced term set of about 200 terms.

During a later stage in the project additional terms were added from the CWR descriptor lists that were published in August 2006. In retrospective, it may have been a mistake to add the descriptor list terms at this stage, because of the

implications this had on creating the ontology structure.

The terms are entered into the ontology in singular form, with plural forms as variants where applicable.

Taxonomic and geographic names and other proper nouns are not included, as these generally represent instances of classes rather than classes and should not be part of an ontology structure.

Ontology

An ontology structure requires the included terms and concepts to be connected in vertical and horizontal relationships. The vertical relationships, the basic hierarchy, is supposed to be based on the subclass concept, where all terms directly or indirectly are linked to the top concept, *Thing*. A lower level term should always satisfy the '*is a*' relation to a higher term. The horizontal relations between the terms are based on other properties. The most common property is a partitive '*part of*' relation.

Because the existing vocabulary databases used as references in this and the preceding project (Agrovoc and CABI) in most cases have used 'part of' relationships in their hierarchy, much work had to be done in redefining the hierarchy to conform to *is_a* relations. E.g. a term like *disease transmission* is defined as a subclass of *pathology* in Agrovoc. In an ontology '*disease transmission*' *is_a* '*biological dispersal*', which in turn is a '*ecological phenomenon*', while '*pathology*' *is_a* '*biological science*' belonging to the *human activity* namespace.

The example above also illustrates the use of *namespaces* in the ontology. *Pathology* belongs to the namespace *human activity* – removing pathology, a science, from the face of the earth would not remove diseases, which belong to the *ecology* namespace.

Using namespaces in an ontology is strictly not necessary, but helps during the construction phase in several ways:

- each term associated with a term on higher level (vertical *is_a*) in the SQL database was also associated with a namespace code. Originally the namespace codes were crude indications of the type of source the term was taken from (geography, agriculture etc) but during the CWR work namespaces were refined and adapted to other existing ontologies.
- one effect the namespaces was that instead of one large hierarchy of associated terms and concepts there was a number of smaller hierarchies that could be processed individually. Work can be concentrated on one particular namespace at a time, or by assigning different teams of experts to work on different namespaces. In the ontology client, e.g. Protégé, namespaces can be imported individually, be dropped and re-imported without rebuilding the whole ontology.
- some namespaces, e.g. 'units', 'numerics', 'time' are more or less universal and immutable across existing ontologies. In a thematic project like CWR it should not be necessary to redefine measurement units since a namespace for measurements has already been built by experts in other projects.
- namespaces helps to resolve ambiguous terms. The cwr project cuts across several scientific fields and it is difficult to avoid terms that have different definitions in different fields. E.g. 'area' is a term commonly used in environmental protection, but 'area' is also a mathematical term with a different definition. Keeping ambiguous terms in different namespaces allows the use of the terms themselves as unique IDS in the ontology.
- in the ontology client the hierarchy becomes more comprehensive both visually and conceptually as only the top terms of the smaller hierarchies constitute the top of the complete ontology. In the ontology client the namespace for each term is indicated by an associated prefix.
- the main part of the ontology consists only of object and datatype definitions – all terms are added by importing namespaces.

Of course, when replacing complete namespaces care must be taken to avoid dropping terms in the old namespace that have been referenced in horizontal relationships from terms in other namespaces as this would break the integrity of the ontology.

The CWR ontology groups the terms into following namespaces (prefixes in parentheses): biology (b), spatial things (c), ecology (e), phenomena (f), earth (g), human activity (h), numerics (n), material things (m), properties (p), processes (r), substances (s), time (t) and units (u) (Table 1).

The classification of biological diversity has some interesting aspects of the use of namespaces. Most classification systems are human inventions, e.g the classification of weeds and utilitarian plants, ornamental plants, medicinal plants and so on, and go under *human activity*. But the evolutionary relationships between the organisms are not part of human activity and thus fall under the biosphere namespace, even if the science of taxonomy is part of human activity. Thus under biosphere and *organism* we have *prokaryote* with subclasses *archeabacterium* and *bacterium*, and *eukaryote* with subclasses *protist*, *fungus*, *plant*, *animal*. Of course these divisions may change as we learn more about evolutionary relationships, but nevertheless they are not human inventions.

The main part of the work with the ontology was concerned with defining the vertical relations, i.e. finding a suitable superclass for each term. Many 'glue terms' had to

NAMESPACE	DESCRIPTION
biosphere	Most things that directly belongs to biology, e.g. structure and functions
ecology	Things that concern interaction between living things and the environment
earth realm	Geography and the physical earth
human activities	Human society and its activities and effects thereof
material things	Non-living things, sometimes related to human activity
numerics	Numbers. math concepts like arrays, coordinates
phenomena	Phenomena that can be observed that does not directly relate to other namespaces
processes	Physical processes
properties	Physical, chemical or spatial properties
spatial things	Things denoting spatial extent
substances	Chemical and composed substances
time	Time concepts
units	Measurement units

Table 1: Namespaces in the CWR *ontology*

TYPE DEF	FORWARD	REVERSE
partitive		
	has part	part of
	has member	member of
causative		
	causes	caused by
	affects	affected by
essive		
	used to make	made from
	source	derived from
temporal		
	develops into	develops from
	precedes	follows
instrumental		
	grows in	growth environment for
	is means for	performed by means of
objective		
	has author	is author of
predicative		
	has property	property of

Table 2: Examples of horizontal relations, object type definitions, their vertical hierarchy and their forward and reverse forms

be added in this process – terms lacking immediate definitions but necessary for linking to higher levels. The adding of terms from the CWR descriptor lists, mentioned above, considerably slowed this process because of the number of unrelated, odd terms, i.e. professional titles, not directly relevant for CWR.

Horizontal relations

The horizontal relations, or properties, in the ontology are mostly partitive, *part_of*, relations. But there are also other types, e.g. causative, temporal, essive and instrumental property relations. These object relationships among themselves form vertical relationships, e.g. 'member of' is a subclass of the partitive property relation (Table 2).

Only a small part of the required and possible horizontal relations were defined before the import of the term set into the ontology structure. This is because a proper ontology program is needed before the horizontal relations can be defined efficiently. Many properties are actually 'required' and 'necessary and sufficient' for the terms, but to define such properties require proper tools that are only available in an ontology client program.

Also, some terms are *always* linked to others ('*all values from*') while other links are not always true ('*some values from*'). As an example, in the CWR descriptor lists *insects* are mentioned as pollinators of some plants. This means that the term *insect*, a subclass of *animal*, needs to be included in the ontology. But all insects do not perform pollination and neither are all plants pollinated by insects, so the horizontal relationship becomes more complicated, '*insect: performs some pollination*', and from the reverse side: '*pollination: performed_by some insect*, *performed_by some wind*' etc.

Many horizontal relationships like the example above have been defined in the relational CWR database before the import to the ontology structure, but many more remain to be defined with the ontology program where more proper tools are provided.

A very complicated example where the horizontal relations could not defined before the import into the ontology client is the *lichen*. A lichen is a fungus, either an ascomycete (most commonly) or a basidiomycete with a component alga and/or cyanobacterium. So the relationship should be something like in Table 3:

Furthermore, the relationships in Table 3 effectively define what a lichen is and thus should be used as a *necessary and sufficient* type of relationship.

If lichen was to be placed as *is_a* subclasses twice, under both ascomycete and basidiomycete, the rule that a subclass cannot belong to two superclasses on the same level would be violated. Thus lichen must be placed directly under *fungus*, on the same level as ascomycete and basid-

LICHEN	
	(has_component some ascomycete or has_component some basidiomycete) and ((has_component some alga and has_component some cyanobacterium) or has_component some alga or has_component some cyanobacterium)

Table 3: Example of complex properties for 'lichen'. The expressions follow the syntax used in Protégé's expression builder.

iomycete.

The examples further illustrate the fact that constructing an ontology is an endless process that requires very specialized knowledge, as well as access to a good ontology client program.

Synonyms, acronyms, abbreviations, sense and spelling variants

Term variants that share definitions should not duplicate the classes in the ontology. Thus a solution had to be found that satisfied this requirement. The ontology has been created in two different formats, DL and Full, depending on which of two possible solutions is preferable.

One solution is to treat alternative terms as datatype properties (Table 4). Through this solution one of the term is selected as the main term while the variants are linked as *literal datatype properties*. Unfortunately the ontology then no longer passes DL validation, because the alternative terms are effectively instances that are treated as classes.

TYPE DEF	FORWARD	REVERSE
synonym		
	has plural	has singular
	narrow synonym	broad synonym
exact synonym		
	has acronym	is acronym for
	has abbreviation	is abbreviation of

Table 4: Examples of term variant handling through literal datatype properties. Vertical relations, as well as forward and reverse forms can be defined as for normal object properties.

The other solution is to use multiple labels for terms that have variants. While not violating DL validation this has the consequence that there no longer is a main, preferred term, but all terms which the same language tag are equal, except for order in which they appear.

The current two main ontology clients available, Protégé and OBO-Edit, have different approaches to solve the problem of term variants. The latter has more elaborate handling of term variants, but at the same time lacks compatibility with RDF/OWL, though conversion utilities exist.

Although not specifically mentioned in the TOR for this project, the results have been delivered in RDF/OWL format. The OBO format would have been another possibility.

Sources and quotes

The CWR ontology uses the Dublin Core extensions to provide sources (URIS), definitions, and last dates of access for the terms included. References to Agrovoc also are provided through the Dublin Core extensions.

Database export, ontology import and beyond

The work on the terms, their definitions and relations was done using a relational database, with the advantage of having possibilities to do semi-automated commands effecting groups of terms. Such operations are not possible after the structure has been exported into an ontology structure for use by an ontology client program, Protégé or OBO-Edit.

The export from SQL to RDF/OWL was done with a script written in Perl that extracted the terms, descriptions, sources from the database, plus vertical hierarchy, term synonyms and other variants, as well as some simple horizontal relationships and produced an import file for Protégé in RDF/OWL format.

After the import to an ontology client the process of changing and redefining becomes a one-by-one process. Adding definitions and sources now largely becomes a matter of cut-and-paste operation. And once extensive changes on relations between terms have been made in the ontology client it will not be possible to go back and produce the import file again from the CWR sources in the SQL database without losing everything that has been done in the ontology program after the import, at least for the namespace concerned.

On the other hand, from this point onward further editing also most certainly needs the assistance of experts in the various fields covered, and thus also becomes a task for a larger team of experts with thorough knowledge in their fields of expertize as well as principles behind ontology structures.

References

[1] Dublin Core Metadata Initiative, 1999. Dublin Core Metadata Element Set, Version 1.1: Reference Description. URL http://www.dublincore.org/documents/1999/07/02/dces/.

[2] Morten Hulden, 2005. Analysis, Examination and Production of a Terminology Set in Relation to Crop Wild Relatives. URL http://fao.untamo.net/cwr/cwr_report1.pdf.

[3] NASA, 2007. Semantic Web for Earth and Environmental Terminology (SWEET). URL http://sweet.jpl.nasa.gov/ontology/.

[4] Plant Ontology Consortium (POC), 2007. Plant Ontology (PO). URL http://www.plantontology.org/.

[5] Stanford University, 2007. The Protégé Ontology Editor and Knowledge Acquisition System. URL http://protege.stanford.edu/.

[6] The Gene Ontology Consortium, 2007. Gene Ontology (GO). URL http://www.geneontology.org/.

[7] The Gene Ontology Consortium, 2007. OBO-Edit. URL https://sourceforge.net/project/showfiles.php?group_id=36855&package_id=192411.

[8] W3C, 2004. OWL Web Ontology Language Semantics and Abstract Syntax. URL http://www.w3.org/TR/2004/REC-owl-semantics-20040210/.