From AGROVOC to the Agricultural Ontology Service / Concept Server An OWL model for managing ontologies in the agricultural domain

Anita C. Liang[1], Boris Lauser[2], Margherita Sini[2], Johannes Keizer[2], Stephen Katz[2]

[1] Perot Systems, Inc.,
acliang@alum.mit.edu

[2] Food and Agriculture Organization of the United Nations,
00153 Rome, Italy,
{boris.lauser, margherita.sini, johannes.keizer, stephen.katz }@fao.org

Abstract. The Food and Agriculture Organization is developing an underlying model for a new knowledge organization system, the Agricultural Ontology Service Concept Server (AOS/CS), using the Web Ontology Language OWL. In this paper, we describe the purpose of the conversion of the AGROVOC thesaurus to this new system, and highlight in particular the core features of the developed OWL model. We go on to explain how it evolves and differs from the traditional thesaurus approach.

Keywords: Ontologies, Thesauri, Semantic Web, OWL, AGROVOC, AOS.

1 Background and Introduction

Since 2003, the Food and Agriculture Organization (FAO) has been concerned with developing a new model for the AGROVOC thesaurus that accounts for semantic and lexical relations in more refined and precise ways. The objective is to build a multilingual repository of concepts in the agricultural domain, the Concept Server (CS). The CS will serve as a base repository from where to build domain specific ontologies and export traditional thesauri, as well as other forms of knowledge organization systems (KOS)[1].

The OWL Web Ontology Language (OWL) [1] has been chosen for representing the model of the Concept Server for several reasons:
First, using an established standard like OWL will provide for maximal interoperability with other systems. Second, an established standard XML/RDF-

[1] KOS are knowledge structures, including authority files, classification systems, concept spaces, dictionaries, controlled lists, taxonomies, gazetteers, glossaries, ontologies, subject heading sets, thesauri, etc.

based format such as OWL is already interoperable with any RDF triple-store, which allows for easy integration of other RDF/XML-based data sources at the storage level and straightforward data processing and visualization. Third, using OWL, ontologies can be shared easily across the Web, since OWL is explicitly able to draw equivalences between classes and individuals across terminologies. Consistency checks can be performed on linked ontologies to identify and resolve conflicts between the ontologies and reasoning can be used to arrive at conclusions beyond those asserted. Fourth, minimized training efforts are required: it is sufficient to refer to publicly available OWL documentation, instead of having to create heaps of documentation for a proprietary system. Finally, having attained the status of a W3C recommendation means that it has become a stable specification that has achieved a high level of technical quality.

Based on these considerations, we developed a new model in OWL that will serve as a skeleton for building ontologies in the agricultural domain.

2 The multilingual issue

To prepare AGROVOC for use as an ontology [2], it is essential to represent concepts by minimizing bias towards a given language or family of languages. That is, to the extent possible, meaning is considered independently of its realization in a particular language. Some domain concepts have lexicalizations in one language but not in others. Additionally, there are lexical relationships like lexical equivalences (e.g., translations, synonyms) that should be captured in an accurate domain model.
The three levels of representation that we are aiming to express in this model are

- concepts (the abstract meaning), for example 'rice' in the sense of a plant,
- terms (language-specific lexical forms), for example 'Rice', 'Riz', 'Arroz', '稻米', 'ข้าว', or 'Paddy',
- term variants (the range of forms that can occur for each term), for example 'O. sativa' or 'Oryza Sativa', 'Organization' or 'Organisation'.

The abstract concepts build the actual hierarchy and semantic structure of the ontology. Terms are no longer arranged in a hierarchy or related via semantic relationships, as is currently done in AGROVOC. Each term is a separate entity in every language that can be linked to concepts, to other terms and to its term variants.

These distinctions allow us to posit the following inter-level relations:

Concept to Term	has_lexicalization (links concepts to their lexical realizations);
Term to String[2]	has_acronym, has_spelling_variant, has_abbreviation link terms to term form variants

[2] String here simply means that the term variants do not constitute a new term, but are simply variant strings of the same term.

Intra-level relations occur at both the level of the concept and at the level of the term exemplified by the following:

Concept to Concept	is_a (hierarchy), pest_of, pest, etc.
Term to Term	is_synonym_of, is_translation_of, is_scientific_term_of

2.1 The OWL model

The design of the new model is done in the OWL DL version of the language, in order to maintain the characteristic of computational completeness.

The baseline of the new OWL model has three concepts at the top level: "category", "classification scheme", "lexicalization". The "category" concept subsumes the concept "domain_concept" (which implies that every domain concept is also potentially a category).

The concept domain_concept is the root of all domain concepts that constitute the core hierarchical structure of the Concept Server. This node subsumes all the basic structural characteristics of the domain ontology, i.e. a hierarchy of classes and their instances along with their relations, axioms, constraints and annotations. Basically, all AGROVOC terms will be modeled under this node.

The separate class category accounts for the need of specific categories that are not domain concepts. Categories are organized in Classification Schemes represented by the class "classification_scheme".

All lexical information is subsumed by the concept lexicalization. Each term that describes a concept in a specific language is modeled as an instance of this concept. Terms are related to the concept whose meaning they lexicalize via two OWL object properties, *has_lexicalization* and its inverse relationship, *means*.

All concept-to-concept relationships (like "is part of", "is infected by", etc.), and the term-to-term relationships (like "has synonym", "has translation", etc.) are organized hierarchically [3]. They are modeled as OWL object properties and the domain and range of all these relationships are set to the corresponding concepts.

Terms themselves can be represented in varying ways: we can have acronyms, shortened forms, abbreviations, etc. A given term is related to its variants through OWL data type properties.

Concept annotations are additional information about domain concepts or categories, some of them coming from the traditional thesaurus world like definitions, comments, scope notes, images and history notes. The model contains these concept annotations as separate instances linked to "category" or "domain_concept". It furthermore contains simple annotations like date created and last updated, status and source (i.e. where the concept has been taken from).

2.2 Backward compatibility

One of the concerns in moving to a new system is compatibility with current legacy systems that will continue to use the traditional thesaurus. We have therefore included further annotations into our model in order to provide full backward compatibility for extracting the traditional AGROVOC thesaurus as it is used today.

3 Roadmap: Where to go from here?

Having completed the basic OWL model for the AOS Concept Server, terminologists need to work on the actual content and the semantics. We are developing a web based maintenance tool, the AOS Concept Server Workbench, which can be used by experts and terminologists worldwide in order to perform the refinement and maintenance work. This tool will be specifically developed for the purpose of editing the complex terminological and conceptual structures modeled in the AOS Concept Server.

4 Conclusion and related work

The OWL model presented in this paper as well as the future AOS Workbench will be open source and we encourage terminology developers worldwide to use the OWL model for representing their KOS and terminology systems. It is true that there are other existing standards and proposals for terminology systems and thesauri. TermBase eXchange (TBX) [4] is an ISO standard for representing terminologies in XML. The Simple Knowledge Organization System (SKOS) [5] format is a W3C proposal for representing simple KOS like thesauri. Our model is different in that it combines new and emerging technologies of the semantic web with the traditional library world of terminology systems and thesauri. Our model subsumes the other mentioned approaches, i.e. we will provide means to create TBX or SKOS compliant extractions. However, our model offers more. It is possible to model more complex ontological structures that can be used in more sophisticated systems.

References

1. OWL Web Ontology Language Reference, 2004. http://www.w3.org/TR/owl-ref
2. Dagober Soergel, Boris Lauser, Anita Liang, Johannes Keizer, Stephen Katz, Frehiwot Fisseha, Reengineering Thesauri for New Applications: the AGROVOC Example, (2004)
3. Proposed Concept Server relationships: http://www.fao.org/aims/cs_relationships.htm
4. TBX web site: http://www.lisa.org/standards/tbx/
5. SKOS web site: http://www.w3.org/2004/02/skos/