Bridging End Users' Terms and AGROVOC Concept Server Vocabularies

Ahsan Morshed[1] Gudrun Johannsen[1] Johannes Keizer[1] Marcia Lei Zeng[2]
FAO[3], Rome , Italy FAO, Rome, Italy FAO, Rome, Italy Kent State University, USA

Keywords: vocabulary; Synonym rings; AGROVOC; concept server

Abstract :

The AGROVOC is a multilingual structured thesaurus in the agricultural domain. It has already been mapped with several vocabularies, for example AGROVOC-CAT, AGROVOC-NALT , AGROVOC-SWD. Although these vocabularies already contained a good portion of non-preferred terms, the terms are collected under the literary warrant and institutional warrant principles; which means vocabularies were collected based on the documents and publications rather than user's queries. It is still very common that end users would use different terms to express the same concept. In light of above discussion, we need to bridge these vocabularies and the users' terms

Background :

AGROVOC is one of the most important resources for covering the terminology of all subjects of interest to the Food and Agriculture Organization of the United Nations (FAO) (including agriculture, forestry, fisheries, food and related domains). AGROVOC is a multilingual thesaurus developed by FAO and the Commission of the European Communities in the early 80s. Since then it has continuously been updated by FAO in collaboration with partner organizations in different countries, and is now available online in 19 languages .

AGROVOC is currently being converted from a traditional term-based knowledge organization system (KOS) to a concept-based system (Soergel, 2004), the AGROVOC Concept Server (CS). The CS allows the representation of more semantics such as specific relationships between concepts as well as relationships between their multilingual lexicalizations. Its functions include being a resource to help structure and standardize agricultural terminology in multiple languages for use by any number of different users and systems around the world. An enabling tool, the AGROVOC Concept Server Workbench (ACSW), has been developed by FAO in collaboration with Kasetsart University in Thailand and other partners. It supports the maintenance of the CS data in a distributed environment (Sini, 2008). One of the goals of the project is to set up a network of international experts who can share the collaborative maintenance and extension of the AGROVOC CS, and thus enhance the creation of agricultural knowledge much more efficiently. The ACSW is part of the larger Agricultural Ontology Service (AOS) initiative and the first major step towards an "Ontology Service" (Fisseha, 2001), which aims to provide semantic-based services to users in the agricultural domain. To cover all agricultural related information, the ACSW needs integrated vocabularies.

Objectives of the Current Sub-Project:

AGROVOC has already been mapped with several vocabularies, for example, *AGROVOC-CAT*[4], *AGROVOC-NALT*[5], *AGROVOC-SWD*[6]. Two other projects are currently ongoing (*AGROVOC-*

[1] Firstname.Lastname@fao.org
[2] mzeng@kent.edu
[3] The Food and Agricultural Organization of UN (FAO)
[4] Chinese Agriculture Thesaurus (CAT)

CAB Thesaurus[7] *and AGROVOC-AgroXML*[8]). Although these vocabularies already contained a good portion of non-preferred terms, the terms are collected under the literary warrant and institutional warrant principles; which means vocabularies were collected based on the documents and publications rather than user's queries. It is still very common that end users would use different terms to express the same concept. In light of above discussion, we need to bridge these vocabularies and the users' terms. The problem is to decide by the system which users' terms should be accepted and mapped with the integrated vocabularies at the back-end of the ACSW. Ideally we should be able to return the results based on the same concept for any given term used to express this concept. Thus the objective of this sub-project is to map the terms not covered by the ACSW vocabularies so that users can navigate all agricultural information and get desired documents when they use their own terms in searching.

Approach and Method of the Current Sub-Project:

Our approach is to introduce synonym rings and map the rings to the entries in the existing ACSW vocabularies. A synonym ring is considered a type of controlled vocabulary and has been written into the ANSI/NISO Z39.50 standard, *Guidelines for the Construction, Format, and Management of Monolingual Controlled Vocabularies* (2005). Unlike other types of KOS which are used during the indexing process, synonym rings are used only during retrieval. A synonym ring is a set of terms that are considered equivalent for the purposes of retrieval. When a concept is described by multiple synonymous or quasi-synonymous terms, a synonym ring ensures that a set of documents will be retrieved as long as any one of the terms is used in a search.For example, a search for the root vegetable "arracacha" should be able to retrieve a set of documents that are indexed with "arracha" as well as with "Peruvian carrot", "Peruvian parsnip", "Arracacia xanthorrhiza" or , "Arracacia esculenta" while there is no requirement for picking one of them as the "preferred" term in searching. Rings can include all kinds of synonyms: true synonyms, misspellings, predecessors, abbreviations, near synonyms, etc. (Zeng, 2008). Synonym rings usually occur as sets of flat lists. The advantages of adding synonym rings at the retrieval process are brought by its dynamic and post-control characteristics. A synonym ring can be built without a fixed file, can be built-on-demand (i.e., to react to the high activities on certain topics overnight), and can be built and executed individually (i.e., no need for a large vocabulary updating project).

Creating synonym rings involves going through word stocks in different sources. One of the main sources for hot topics (an example would be 2009 H1N1 Flu related terms at the hot peek period) is the search logs. Other progressively built rings would follow the evidence of the FAO search traffic and priorities of FAO units. A selected list of sources of the terms consists of: search logs, dictionaries, metadata content embedded in Webpages, terms from dbpedia under certain properties (such as owl:sameAs (link), rdfs:lable, skos:prefLabel, skos:subject, dbpprop:name, dbpprop:disambiguates, dbpprop:redirect, etc.) and other KOS schemes. After testing by humans, automatic programs can be written to automatically collect candidate terms and filter terms according to the testing on particular source (e.g., for dbpedia's individual property). The testing process helps to decide what terms should be considered interchangeable when searching.

Terms that are considered to form a synonym ring can be stored as a unit in a search system. An example are terms used around 'ship' (including 'vessel', 'boat', 'sailboat', etc.) and the linguistic variants of these terms (over 40 multilingual terms collected automatically). A search using any term in the ring will retrieve all documents tagged as designated. In the following

[5] National Agricultural Library Thesaurus (NALT)
[6] Schlagwortnormdatei
[7] http://www.cabi.org/cabthesaurus/
[8] http://www.agroxml.de/

Figure, the term "vessel" is entered into the search box, it maps to only one URL (which represents the concept resulted through the mapping processes in the Concept Server) and the system will return all documents about or related to this concept.

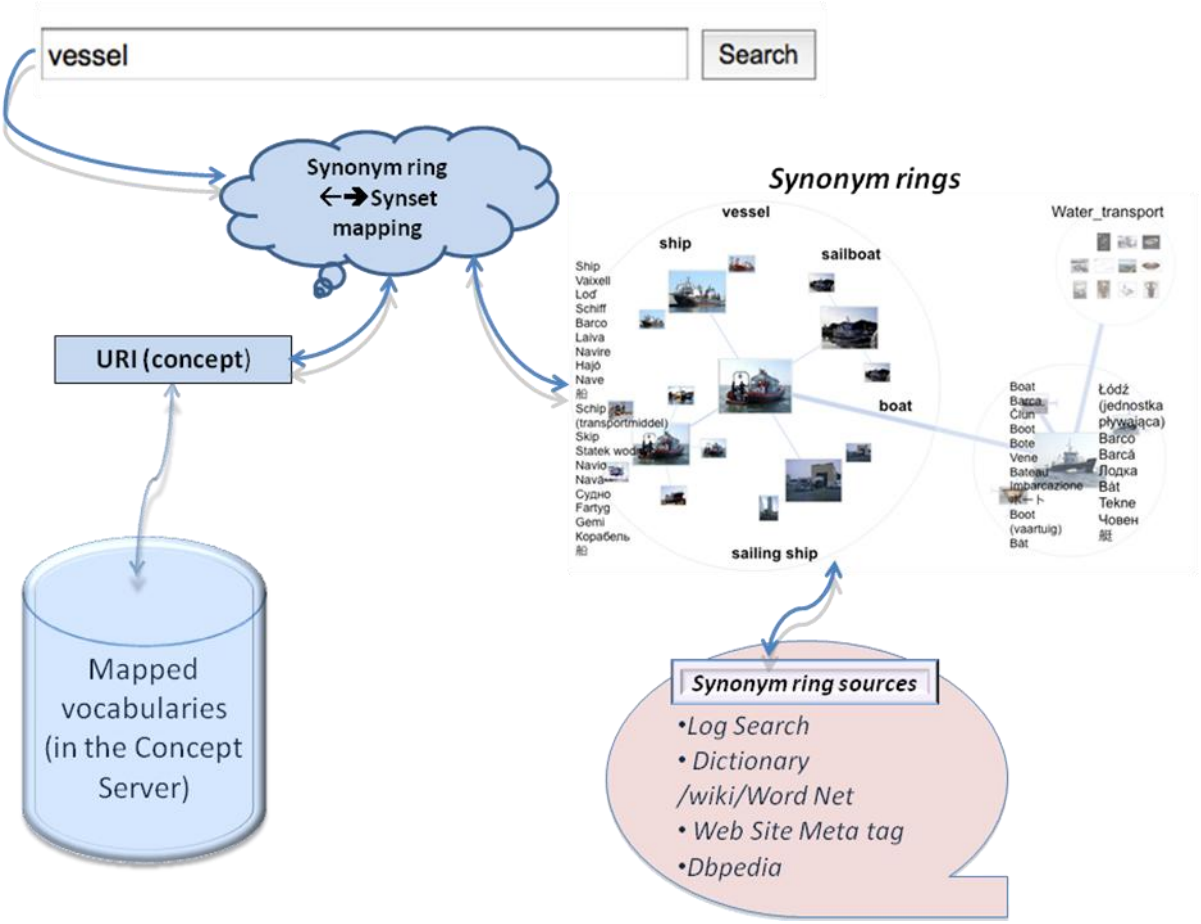

FIG 1:Illustration of the Search System Supported by Mapping Synonyms and ACSW Vocabularies.

Because users can be confused by results that do not actually include their keywords, interface design and an understanding of user goals become the keys for proper balance. A search interface may provide a clue about what terms are considered synonyms. Google and other search services' spelling-based substitute search (i.e., "Did you mean X?" for a single term or "Showing results for X Y Z. Search instead for X Y A" for a phrase) are considered as the interface of this layer.

Conclusion: This sub-project is part of our ongoing AGROVOC Concept Server Workbench project, while turning the focus on mapping users search terms to internal vocabularies. Adding synonym rings as a bridge supporting the traffic between the search box and the ACSW mapped vocabularies can overcome the barriers of unmatched search queries with the controlled vocabularies. It will give particular information about the search terms and enable better control in the search browser. This poster will present the illustration of the architecture behind this new service and experiment results. It will also report the testing of automatically collected synonyms from beyond query logs. The exploration of reusing linked data is a new and applicable methodology. Lessons learned as well as the methodology can be applied by others who have similar goals.

References

ANSI/NISO Z39.19 - *Guidelines for the Construction, Format, and Management of Monolingual Controlled Vocabularies*. (2005). < http://www.niso.org/kst/reports/standards/>

Fisseha, Frehiwot, Hagedorn, Kat, Keizer, Johannes, and Katz, Stephen, (2001), "Creating the Semantic Web: the Role of an Agricultural Ontology Server (AOS)". *International Semantic Web Workshop (Infrastructure and Applications for the Semantic Web*. <ftp://ftp.fao.org/docrep/fao/008/af228e/af228e00.pdf>

Lauser, Boris, Margherita Sini, Anita Liang, Johannes Keizer, and Stephen Katz , (2006), "From AGROVOC to the Agricultural Ontology Service / Concept Server An OWL model for creating ontologies in the agricultural domain". *Proceedings of the OWLED*06 Workshop on OWL: Experiences and Directions*, Athens, Georgia, USA, November 10-11, 2006. <ftp://ftp.fao.org/docrep/fao/009/ah801e/ah801e00.pdf>

Sini, Margherita, Boris Lauser, Gauri Salokhe, Johannes Keizer, and Stephen Katz, (2008). "The AGROVOC Concept Server: rationale, goals and usage". *Library Review*, 57(3):200 – 212. < http://agrovoc-cs-workbench.googlecode.com/files/Final-322881-1.doc>

Soergel, Dagobert, Lauser, Boris, Liang, Anita, Fisseha, Frehiwot, Keizer, Johannes, and Katz, Stephen. (2004), "Reengineering thesauri for new applications: the AGROVOC example", *Journal of Digital Information*, 4(4). <ftp://ftp.fao.org/docrep/fao/008/af234e/af234e00.pdf>and
< http://journals.tdl.org/jodi/article/viewArticle/112/111>

Zeng, Marcia L. (2008) Knowledge Organization Systems. *Knowledge Organization*. 35(2-3): 160-182.