The Mapping Schema from Chinese Agricultural Thesaurus to AGROVOC

A. Liang [a], M. Sini [a]
Chang Chun [b], Li Sijing, Lu Wenlin [b], He Chunpei [b]
J. Keizer [a]

[a] *Library and Documentation Systems Division, Food and Agriculture Organization of the United Nations (FAO), Rome, Italy, johannes.keizer@fao.org*

[b] *Agricultural Information Institute Chinese Academy of Agricultural Sciences (CAAS), Beijing China, changc@mail.caas.net.cn, lsj@mail.caas.net.cn*

Abstract

This paper introduces the criteria and the procedures for mapping the Chinese Agricultural Thesaurus (CAT) to FAO's multilingual agricultural thesaurus AGROVOC. It proposes modifications to the inter-thesaurus mapping rules provided in the Simple Knowledge Organization System (SKOS) specification. It outlines in detail the criteria for the application of each of the mapping rules. We will describe the procedure for the application for these rules and give concrete examples taken from both thesauri.

Key words: Mapping, Thesaurus, CAT, AGROVOC, SKOS.

1 Introduction

The objective of this project is to link two multilingual terminology sources related to agriculture, the Chinese Agricultural Thesaurus[1] (CAT) and AGROVOC[2]. The motivation is to create correspondences between different world views/multilingual perspectives, to mutually enrich both terminologies in terms of domain and language coverage, and to improve their structure. Because mapping increases the interoperability of the source terminologies, information retrieval applications that include functionalities such as cross-linguistic searches and terminology brokering will be made possible (see fig. 1).

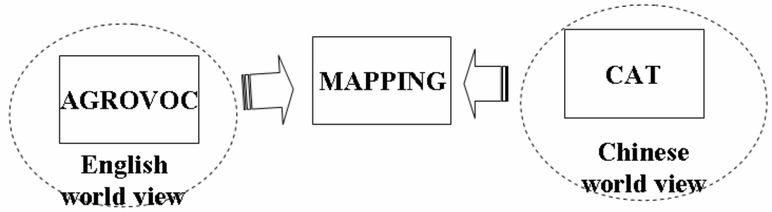

Fig. 1: Reconciling different world views

In addition, in terms of FAO's Agricultural Ontology Service (AOS) initiative, this project will demonstrate the feasibility of incorporating other terminologies within the AGROVOC Concept Server via mapping. By mapping we mean the creation of an explicit link, using our specifications of the Simple Knowledge Organization System[3] (SKOS) mapping rules, from one source (CAT) to another (AGROVOC). Source and target vocabularies may be revised, when inconsistencies or error are encountered, but each is structurally independent of the other; mutual consistency is desirable but less a priority than establishing approximate equivalences.

A test has already been carried out on some sample data involving top terms and their lower branches (child nodes) in the cereal crops sub-domain from both CAT and AGROVOC.

2 The structure of CAT and AGROVOC

Given that CAT plays an important role in agricultural information management in China and AGROVOC is used widely to index agricultural information material all over the world, if CAT is mapped to AGROVOC, then Chinese agricultural scientists will be able to access resources indexed with AGROVOC using CAT terms, and scientists outside of China can access CAT resources using AGROVOC.

CAT was developed as a knowledge management tool for the agricultural forestry and biological fields. It is the second largest multi-disciplinary thesaurus in China. It has been approved by the authority of the Agricultural Ministry of China as a criterion for Agricultural document retrieval systems and for archiving administration and scientific research resources. CAT has been extended in Taiwan by the Taiwan Agricultural Science Information Center. CAT contains 64638 terms in Chinese, including 51614 descriptors and 13024 non-descriptors. It has BT/NT, UF/USE and RT relations. Most descriptors have English translations. Some biological taxonomic names have only Latin translations. Only 200 descriptors and nearly all non-descriptors have no translation. There are 2332 top terms and terms are organized in 40 categories and sub-categories (indicated by codes 01, 19, 50, etc., e.g. crops) up to a maximum depth of three.

FAO published the first edition of AGROVOC in 1982, and then issued the second, third and fourth editions in 1988, 1995, and 1999, respectively. FAO issued the Web edition in 2000 which now has seven languages (Arabic, Chinese, English, French, Spanish, Czech, and Portuguese). AGROVOC has a total number of 16769 descriptors and 10968 non-descriptors in the area of agricultural science. The Chinese translations were provided by the same thesaurus experts who manage CAT [5]. Initially, AGROVOC was used for indexing information materials produced within the international cooperative information systems AGRIS and CARIS, and for data retrieval from those systems. Nowadays, AGROVOC is used more widely to index agricultural information in repositories all over the world.

2.1 Structural differences to consider when mapping

Not only does CAT have more than twice as many terms as AGROVOC, but it covers a broader range and goes more deeply than AGROVOC does in the Agricultural sub-domains. For some domains, whereas AGROVOC has up to 4 levels of narrower terms, CAT has up to seven.

3 The Mapping Schema

As mentioned in the previous section, the source vocabulary is CAT and the target vocabulary is AGROVOC. Mapping means linking an entry in the source vocabulary to an entry in the target vocabulary. An entry in CAT consists of the Chinese term and any English translation(s) along with its relations to other entries. An entry in AGROVOC consists of at least one English or Chinese term along with their translations as well as its relations to other entries. We use the term 'concept' and 'entry' interchangeably. A term is a lexical representation of a concept. Note that entries do not necessarily have to have lexicalizations in both Chinese and English.

Therefore, in order to carry out a mapping between two concepts, both Chinese and English lexicalizations, when they occur in a given entry, must be considered. The following example demonstrates why: CAT '水稻' / 'Oryza sativa' was originally mapped to AGROVOC 'Oryza sativa'. However, upon closer examination, the Chinese lexicalization in AGROVOC of 'Oryza sativa', which is '稻', appears to be the broader term of the CAT Chinese term. Moreover, a search in AGROVOC for the CAT Chinese term '水稻', shows the English translation as 'Paddy'. These discrepancies indicate the weakness of the mentioned procedure and the necessity of cross checking all lexicalizations in both languages.

The relationships are drawn from the SKOS Mapping Vocabulary Specification[4] (version 2004), which define the characteristics of each of the following properties:

- exactMatch
- broadMatch
- narrowMatch
- majorMatch
- minorMatch

and the following classes:

- AND
- OR
- NOT

We tried to apply the rules to sample data according to the SKOS specifications, but we found that it was difficult for several reasons. First, the SKOS rules assume that the thesauri to be mapped have both been used to index the same set of resources, which is not the case with the current project. Secondly, the SKOS descriptions of the mapping properties are not well defined, especially for the problems of thesauri in unrelated languages, such as Chinese and English. Third, the heterogeneity of the terminologies themselves complicates the work. Therefore, we modify the rules in such a way as to be able to perform the mappings between the thesauri given the particular linguistic and conceptual issues that characterize these terminologies. We also assume that our modifications are applicable to other projects involving mapping of multilingual thesauri.

3.1 Procedure

Prior to applying the mapping rules, a set of guidelines should be kept in mind:

1. Entries should be mapped irrespective of their status as descriptors or non-descriptors;
2. Mappings should be between entries, not terms;
3. Many to one: more CAT entries could be mapped to the same entry in the target vocabulary;
4. One to many: an entry in CAT could be mapped to one or more entries in the target vocabulary.

Then, the overall mapping process will be performed according to the following sequence:

1. find the exact match at the highest possible level;
2. if there is no exact match, find an approximate equivalent and apply the broad / narrow match;
3. map the corresponding children of each mapped concept;
4. check inheritance of all the non mapped children.

3.2 Inheritance

In cases where there are descendents of a concept in the source vocabulary and there is a mapping from that concept to one in the target vocabulary, and there are no corresponding children in the target vocabulary, rather than mapping each individual descendent via a broad match to the target concept, we assume that those descendents are implicitly map by inheritance.

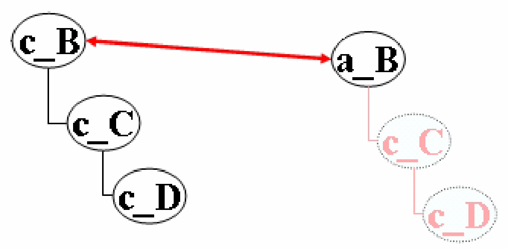

Fig. 2: The inheritance mechanism.

For example, CAT ‘数学’ / ‘Mathematics’, corresponds to AGROVOC ‘数学’ / ‘Mathematics’. Furthermore, the CAT concept has over 200 descendants whereas the AGROVOC concept includes none of them. Instead of mapping each individual descendant to the AGROVOC concept using a broad match mapping rule, inheritance is applied. This means that the mapping file will not include any reference to the descendants of the mapped concept unless those children have corresponding equivalences in the target vocabulary.

3.3 ExactMatch

We consider concepts to be the same if:

- they have the same Chinese and English terms in both thesauri;
- they have the same Chinese terms and English terms are synonyms;
- they have the same English terms and Chinese terms are synonyms;

even if they have different BT/NT/RT, e.g.

123: CAT ‘禾谷类作物’ / ‘Cereal crops’.

2551: AGROVOC ‘Cereal crops’ / ‘禾谷类作物’.

Therefore the mapping would be: CAT-123 *ExactMatch* AGROVOC-2551

When a gap occurs in either vocabulary because the corresponding term is missing, the term should be added to the appropriate vocabulary

3.4 broadMatch and narrowMatch

When a gap occurs in the target vocabulary because the concept does not exist, but there is a concept that is closely related, the broadMatch or narrowMatch property should be used. If the target concept is more general, then the broadMatch should be used. If it is more specific, then the narrowMatch should be used. See Figure 2.

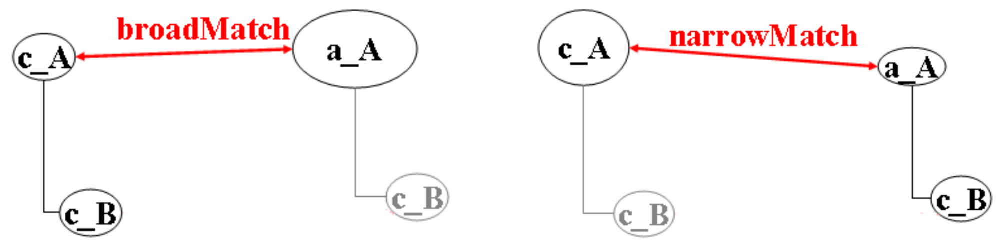

Fig. 3: The broadMatch and narrowMatch relationships, including inheritance (see below).

By inheritance, c_B is implicitly mapped to a_A as a narrow term, and vice versa.

3.5 partialMatch

SKOS suggests using majorMatch or minorMatch to indicate the mapping relationship between two concepts between which there is some degree of semantic overlap. However, the definitions of major and minor match are imprecise, and in practice, they are difficult to use as specified. So, we redefine one of the rules outlined in the SKOS specifications. SKOS defines a mapping rule called partialMatch but because it is supposed to represent a subsumption relation, we find this to be a misnomer; therefore, we redefine this term to mean the link between any two concepts that have some degree of overlap in their meaning excluding subsumption, which is already covered by the broad and narrow match.

For example, the CAT concept ‘经济大国’ / ‘Economic power’ has been analyzed to be a partial match with the AGROVOC ‘Developed countries’ / ‘发达国家’ because not all developed countries are sub-concepts of ‘Economic power’.

3.6 AND, OR and NOT classes

Because SKOS recognizes that one-to-one correspondences do not always exist between languages and/or terminologies, it introduces the AND, OR, and NOT classes for combining or excluding concepts. ‘AND’ is used to identify a concept formed from the intersection of two or more concepts. ‘OR’ represents the union of the semantics of two or more concepts. ‘NOT’ can be used to create a mapping target from which one or more elements of meaning are excluded.

CAT ‘防火机构’ / ‘Fire control organization’, corresponds to AGROVOC ‘Fire control’ / ‘火灾控制’ AND ‘Organization’ AND ‘Public services’ / ‘公用事业’. Note that, as in the case of AGROVOC ‘Organization’, it is not necessary for there to be a Chinese lexicalization for a concept to be mapped to.

In Chinese there are different terms (concepts) used to lexicalize the grain and the plant. The term ‘大米’ is used as rice as a grain which can be eaten; whereas ‘稻’ is used for the plant ‘Oryza sativa’. Normally however, the distinction between the crops plant and the grain, is not made. For instance the CAT ‘大麦’ is used for both the plant and the grain. In such cases, we use the rule of OR to map these terms. Thus, in the case of barley, the CAT term ‘大麦’ is exact matched with AGROVOC ‘Hordeum vulgare’ / ‘大麦(植物)’ OR ‘Barley’ / ‘大麦’[1].

CAT ‘大陆’ / ‘Continent or Mainland’ is a geographic term, indicating the part of China excluding Hongkong, and Macau. So we can use NOT rule to map this term, ‘大陆’ exact match with ‘China’ NOT (‘Hongkong’ AND ‘Macau’).

3.7 Special cases

During our preliminary test phase we encountered the following issues:

1) CAT concept ‘国外’ / ‘Abroad’ - ‘External’ - ‘Overseas’ does not exist in AGROVOC. Therefore it is suggested that it be added to the target vocabulary.

2) Some special cases that may occur within CAT or other multilingual mapping projects, are currently unresolved using the described formalism: for example the French concept ‘tutoiement’ is difficult to represent in a language that does not distinguish between format and informal uses of the pronoun ‘you’. Feedback or comments are welcome for this situation.

4 Discussion and Conclusions

For this project we propose the use of a subset of the SKOS rules modified to deal with the semantic heterogeneity stemming form differences in language terminology domain coverage granularity. We provided some examples and guidelines on how to apply the rules. In addition we also introduced the notion of inheritance.

This work is just at the beginning. At this time, we can concentrate more on the design of the mapping procedure and the testing phases. As we perform the work we are likely to encounter other unanticipated problems and issues that will require revision of our initial guidelines. We also believe that this work can benefit other terminology mapping projects especially by contributing to the discussion concerning the refinement of the SKOS specification based on its application to real world data.

[1] Due to the way that AGROVOC is structured, the distinction between the term and the concept is only approximate. The lexicalizations of barley should all be part of the same entry. Thus, there should be an exact match with a single barley concept that subsumes both the grain and the plant notions.

As a mapping procedure involves a great number of terms and mapping relations, it is nearly impossible to do the job without the help of computers. For instance automatic matching of CAT and AGROVOC terms with the same English and Chinese can be proposed for exactMatch mappings. For concepts that are not mapped an automatic procedure can be applied to check if an ancestor has been mapped and if not the concept can be flagged for manual mapping. The use of tools such as Protégé and Prompt can also provide semi-automatic assistance on the mapping.

5 References

[1] Chinese Academy of Agricultural Sciences, 1994. Agricultural Thesaurus. Chinese Agricultural Press.

[2] Food and Agricultural Organization of the United Nations. Multilingual Agricultural Thesaurus (AGROVOC). http://www.fao.org/agrovoc/ (Accessed April 16, 2005)

[3] Miles, A., Matthews, B. Inter-Thesaurus Mapping. http://www.w3c.rl.ac.uk/SWAD/deliverables/8.4.html (Accessed April 16, 2005)

[4] Miles, A., Brickley, D. SKOS Mapping Vocabulary Specification. http://www.w3.org/2004/02/skos/mapping/spec/ (Accessed May 16, 2005)

[5] Chang Chun, Lu Wenlin. The translation of agricultural multilingual thesaurus. AFITA2002, Asian Agricultural Information Technology & Management. Proceedings of the Third Asian Conference for Information Technology in Agriculture. Edited by Mei Fangquan. China agricultural scientech press,2002:526-528