Requirements for the treatment of multilinguality in ontologies within FAO

Caterina Caracciolo, Margherita Sini, Johannes Keizer

Food and Agriculture Organization of the United Nations,
v.le Terme di Caracalla 1,
00153 Roma, Italy
{caterina.caracciolo, margherita.sini, johannes.keizer}@fao.org

Abstract. International organizations like FAO are intrinsically multilingual. FAO is currently experimenting with semantic-oriented technologies based on ontologies, with the purpose of integrating data across various information systems and providing better services to end users. However, in order for these technologies to be used in real-life scenarios, models and tools for accommodating and managing multilingual data are needed. This paper analyzes the requirements for the treatment of multilinguality as resulting from the experience we gained at FAO.

Keywords: multilinguality; ontologies.

1 Introduction: multilinguality at FAO

Multilinguality is central to all FAO's activities. Since official documents and resources produced by FAO must be made available in all the five FAO languages (Arabic, Chinese, English, French, and Spanish), all FAO information systems manage multilingual data, be that textual documents, terminological databases, and reference data to access statistical databases. The type of multilinguality associated with these resources is variable. On the one hand, there is the reference data used to access statistical data, where a one-to-one correspondence between names in various languages is firmly established on the basis of agreements and conventions (e.g. names of countries and organizations). On the other hand, there are resources such as the AGROVOC thesaurus, the content of which heavily depends on the culture in which it is developed. All data is stored in relational databases.

Currently, a number of pilot studies are being carried on within FAO to improve the level of interoperability among FAO information systems (IS) by means of semantically oriented technologies. In order for these pilot studies to be successful the multilingual issues need to be carefully addressed: in terms of models, software, support to editing and interfaces to editors.

In the rest of this paper we illustrate in some detail the multilingual resources available at FAO, and the future direction currently under study (Section 2). In Section 3 we provide our requirements for the treatment of multilinguality, and in Section 4 we discuss them.

2 Multilingual resources within FAO

The reference tables used to store and access statistical data are an example of multilingual resources where the one-to-one correspondence between languages is given. For example, the reference tables on fisheries (currently managed by a Reference Table Management System (RTMS) [1]), contains the list of all classification systems and hierarchies used to deliver statistics, and most available fact sheets.

AGROVOC [2], a multilingual thesaurus (i.e., a structured controlled vocabulary) designed to cover the terminology of all subjects of interest to FAO (agriculture, forestry, fisheries, food and related domains such as environment), is an example of a resource where there are loose relationships between terms in various languages. The main role of AGROVOC is to serve as the basis for uniform indexing of textual documents. It was developed by FAO and the Commission of the European Communities in the early 1980s. Since then it has been continuously updated by FAO and new online releases are issued about four times a year. Currently, AGROVOC includes around 35,000 terms per language, counting both descriptors and non-descriptors. The available languages are the five FAO languages, together with Czech, Portuguese, Japanese, Thai, Slovak, Hungarian, Polish, and German. Versions for Italian, Korean, Farsi, Hindi and Lao are currently under development, and possible future versions include Amharic, Catalan, Russian and Moldavian.

The translations of AGROVOC are generally provided by domain experts that are also native speakers of the target language. Translations of AGROVOC are usually made by national Institutions or Ministries (often from the English version) and sent back to FAO for validation and inclusion in the master copy. Therefore, AGROVOC is not completely and "purely" speaking a multilingual thesaurus, but a translated thesaurus [3] [4].

Semantically oriented technologies promise to improve the level of interoperability between IS. Therefore, FAO is experimenting with the possibility of using inter-connected and networked ontologies [5] [6]. Also, AGROVOC is the foundation that underpins the development of the Agricultural Ontology Service (AOS) initiative [7]. By making use of knowledge contained in vocabulary systems and thesauri such as AGROVOC, AOS is committed to developing specialized domain-specific ontologies and concept-based structures that will better support information management for the web environment. During our work we found that a number of issues related to the representation of multilingual information need to be addressed. We list them in the next section.

3 Requirements

When working with multilingual ontologies we should be able to:

1. represent lexicalizations in several languages that refer to the same object; for example ‘cow’ (English), ‘vache’ (French), ‘母牛’ (Chinese), ‘vacca’ (Italian) and ‘mucca’ (Italian) all refer to the same animal [8];
2. specify ISO standard codes for languages and countries for lexical items;
3. represent relationships between lexical items, within and across languages. These relationships include synonymy (e.g. ‘vacca’ and ‘mucca’ are synonyms and both translate the Chinese ‘母牛’), acronyms (‘BSE’ stands for the English ‘Bovine spongiform encephalopathy’, but ‘ESB’ stands for the French name of the same disease, ‘Encéphalopathie spongiforme bovine’), spelling variances, different types of names (e.g. ‘Gabonese Republic’ is the official long name, and ‘Gabon’ is the official short name of the same country);
4. impose constraints at the lexical level, such as the number of allowed translations and acronyms;
5. add additional information at the lexical level, such as comments, definitions, and images, in order to better specify the intended meaning of a term;
6. account for the use of the same terms in different context, that are translated by different terms in other languages (e.g. ‘benzene’ is considered a pollutant in a specific area, and a fuel in another sphere [9]);
7. account for culturally determined relationships that should not be “inherited” by translation. For example the animal ‘scorpion’ can be seen as related to food in some African countries or in China, but not in other parts of the world (therefore not in other languages);
8. manipulate “layers” of languages: to extract one or more languages for the purpose of editing, data export or visualization, and to add new languages;
9. use UTF-8 as default character encoding (in some cases UTF-16 should also be supported);
10. support left-to-right and right-to-left languages.

4 Discussion

Requirement 1 and (partially) Requirement 2 can be met by using the *xml:lang* attribute that allows one to specify the language in which a given element is expressed. However, the simple use of this attribute may not be enough to satisfy Requirements 3, 4 and 5. For instance we are not able to set relationships between terms in different languages that may be used to represent the same object, and we may not be able to assign specific properties to those objects. Requirements 6 and 7 are both related to the concerns expressed in Requirements 1 and 3, i.e. that the linguistic (terminological) level is distinguished by the “factual” level. They also somewhat imply that languages have specificities and that ontologies (and thesauri,

for the matter) built in one language and then translated can lead to counterintuitive results.

Requirement 8 is relevant both to the exploitation of large multilingual ontologies and to their maintenance. In fact, for efficiency reasons it may be very useful to be able to select and extract only one or few languages instead of manipulating the entire data set. Also, during the editorial process it may be very useful to be able to add a new language layer for the entire ontology (as opposed to adding a new element, say a property 'title', individually for each element).

Requirements 9 and 10 address the lowest level features of a multilingual ontology, namely the possibility of properly encoding and visualizing a wider variety of languages than just the western ones.

Most of our requirements can be addressed by ad hoc modeling solutions (e.g. representing the lexical items as specific entities, classes or instances, characterized by an *xml:lang* property, and using object-type relationships to link terms). These ad hoc models can then be accessed and exploited by software components that are bound to be hardly reusable and non-generalizable. We raise the concern that some more general treatments of multilinguality should be provided: be that at the level of ontology language, modeling, or applications for editing and exploiting ontologies.

Finally, one issue to take into consideration for the implementation of the requirements at which we arrived is the existence of massive legacy data (e.g. relational and terminological databases in various formats) that will continue to be used and to which new information systems will need to refer or connect.

References

1. http://www.fao.org/figis/servlet/RefServlet
2. http://www.fao.org/aims/ag_intro.htm
3. M. Doerr, "Semantic Problems of Thesaurus Mapping", *Journal of Digital Information*, Volume 1 Issue 8, 26 March 2001
4. D. Soergel, "Large multilingual vocabularies. Structure and software requirements", Proc. of 61st Annual Meeting of the American Society for Information Science. 1998.
5. C. Caracciolo, M. Iglesias Sucasas, J. Keizer, "Towards Interoperability of Geopolitical Information within FAO". *Computing and Informatics*. Forthcoming.
6. http://www.neon-project.org
7. A. Liang, B. Lauser, M. Sini, J. Keizer, S. Katz, "From AGROVOC to the Agricultural Ontology Service / Concept Server. An OWL model for managing ontologies in the agricultural domain", OWL workshop, 2006, Athens, Georgia, U. S. A. - Proceedings OWL: Experiences and Directions Workshop Series, 2006
8. Soergel, D. Lauser, B. Liang, A. Fisseha, F. Keizer, J. Katz, S., "Reengineering Thesauri for New Applications: the AGROVOC Example". Journal of Digital Information, vol.4, no.4
9. F. Mazzocchi, P. Plini, "Thesaurus classification and relational structure: the EARTh experience", Proc. of 7th International conference on Terminology and Knowledge Engineering (TKE2005).