Multilinguality in AGROVOC Concept Scheme: Challenges and Experiences

Michelle Lim Sien Niu[1], Gudrun Johannsen[2], Ahsan Morshed[2], Sachit Rajbhandari[2], Johannes Keizer[2], Lavanya Kiran[3], Aree Thunkijjanukij[4], Nor Ezam Selan[1]

[1]MIMOS Berhad
Technology Park Malaysia, Kuala Lumpur, Malaysia 57000
Email: {michelle.lim | nor.ezam}@mimos.my

[2]Food and Agriculture Organization of the United Nations
Viale delle Terme di Caracalla. 00153 Rome, Italy
Email: {Gudrun.Johannsen | Ahsan.Morshed | Sachit.Rajbhandari | Johannes.Keizer}@fao.org

[3]National Institute of Plant Health Management (NIPHM)
Rajendranagar, Hyderabad-30
Email: lavanyakirann@gmail.com

[4]Thai National AGRIS Centre
Kasetsart University, Thailand
Email: libarn@ku.ac.th

Abstract

AGROVOC plays a significant role in setting the standards as a common data model for representing and linking multilingual information from Agriculture, Forestry, Fisheries, Food and other related domains resources unequivocally. Researchers, librarians, terminologist and information managers from international organizations leverage on AGROVOC as the controlled vocabulary to access, translate and share their knowledge collaboratively. Since 1982, AGROVOC has evolved from a traditional thesaurus to an ontology based OWL[1] model and finally to a SKOS-XL[2] model. Based on the model, FAO[3] built a web based vocabulary management tool called "VocBench" which can handle the multilinguality concurrently. This paper outlines the challenges of multilinguality in AGROVOC Concept Scheme, as well as the experiences and processes of converting multilingual information from heterogeneous data formats into the AGROVOC Knowledge Base.

Keywords: *AGROVOC, Agriculture, Vocabularies, Thesaurus, Ontology, OWL, Linked Open Data, Knowledge Base*

1. Introduction

AGROVOC [1], developed by FAO and the Commission of the European Communities, is one of the most important knowledge organization systems (KOS) in the fields of agriculture, forestry, fisheries, food, and other related domains. First published in 1982 as a thesaurus in English, French and Spanish, it has been used by libraries and documentation centers for indexing and retrieving agricultural information resources. As users felt the need to index and search information in their own language, AGROVOC has been translated by national agricultural research institutions in different countries, and it currently includes 579,523 terms in 19 different languages. Furthermore, due to the growth of the Semantic Web,

[1] Web Ontology Language (OWL).
[2] Simple Knowledge Organization System (SKOS).
[3] The Food and Agriculture Organization of the United Nation (FAO).

AGROVOC underwent a process of semantic refinement to shape it into a reusable Resource Description Framework (RDF) vocabulary that could fulfill its traditional functions of document indexing while also meeting the needs of a new generation of semantically-enabled applications. FAO made the first port of AGROVOC [2] using a customized model based on Web Ontology Language (OWL).

In 2010, FAO starts to migrate AGROVOC to SKOS format. Fig. 1 depicts the AGROVOC SKOS model, a semantically structured concept-based system which consists of concepts with their lexical representation, specific relationships between concepts and also relationships between their multilingual lexicalizations.

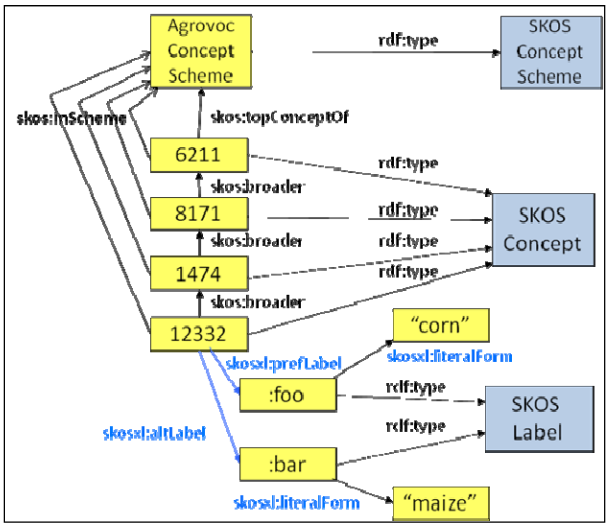

Fig. 1. AGROVOC SKOS model

The model was tailored to represent concept schemes and provides metadata descriptors about edited resources. It also put emphasis on multilinguality and the natural language description of resources. Such a customized model was difficult to manage using traditional ontology editing tools such as Protégé [3,4] (e.g. reified labels represent different resources to be edited, while users expect them to be editable inside a view of the concept to which they share attachment). The poor usability of traditional ontology editing tools combined with the need for a collaborative environment that supported roles-based authentication, editorial workflow, multilingual search and high installability (or, even better, no installation at all) led to the development of VocBench, a web application specifically tailored to the AGROVOC Vocabulary. By using this tool, any authenticated users can add or translate AGROVOC terms in their languages.

This paper presents the SKOS model[5] and experiences from different editors in different languages. Therefore, any new users can influence to integrate their vocabularies with AGROVOC. The rest of this paper is organized as follows. Section 2 illustrates about the legacy data and experiences from different language editors. Section 3 discusses the challenges in multilinguality while Section 4 outlines the future works and additional enhancements that can be done to further contribute to enhance the multilinguality in AGROVOC Concept Scheme.

2. Multilinguality in AGROVOC

AGROVOC is used widely for information exchange and retrieval among different languages around the world. Therefore, it is crucial to translate the thesaurus into as many languages as possible in order to make it easier for the users to index or search for information sources in their own language. National organizations and institutes are invited to translate AGROVOC into their local languages [6].

2.1. AGROVOC Legacy data

AGROVOC has evolved through many changes in its data model. It was initially developed as a traditional thesaurus in different formats such as MySQL, TagText and XML and store in the database system. Apart from that, the AGROVOC was also stored in the printed version since 2000. Previously, many agriculture libraries, and agriculture centers were using the printed copies of the AGROVOC for their application such as cataloging the resources. Since 2002, the AGROVOC is published online for the wider agricultural audiences. However, the most challenging task was to maintain and update the AGROVOC simultaneously while reducing the laborious human tasks. Due to these necessities, FAO has built a distributed system called VocBench for managing the vocabularies and workflows from anywhere in the world. This has resulted in AGROVOC terms to be translated into different languages by partner organizations of FAO. In the following sub-sections, we will describe the experiences of some of these efforts.

2.2. Bahasa Malaysia (Malay) Terms Translation

Malay is a major language of the Austronesian family[4]. MIMOS Berhad started to participate in the Malay terms translation in the year 2011, with the main

[4] Malay is the official language of Malaysia (as Bahasa Malaysia), Indonesia (as Bahasa Indonesia), and as Bahasa Melayu in Brunei, Thailand, Southern Philippines and Singapore [14].

objective of developing a Malay Agriculture thesaurus for the agriculture communities in the country. This effort is carried out by 5 Knowledge Engineers from the Knowledge Technology Cluster in MIMOS Berhad. The process of Malay language terms submission is performed online through the VocBench itself. As of September 2011, there are a total of 636 Malay terms which has been translated and published in the VocBench.

There are two groups of user roles that serve in the contribution of the Malay terms translation effort. They are term editors and the publisher (who also has the rights as a validator). The term editors are responsible for proposing and managing the concepts and relationships in Malay. While the publishers are responsible for validating (approve or reject the proposed Malay terms submitted by the term editors) and publishing the validated entities. Pusat Rujukan Persuratan Melayu @ DBP Malaysia [7] and Mykamus Mobile [8] are the reference tools used in assisting with the Malay terms translation.

The process of translation starts with the user (term editor) logging into the VocBench and begins searching for the preferred English term that they want to translate. For example, if the user would like to submit the Malay term of "Maize" as "Jagung", the user will first need to search and select the preferred English term "Maize" as depicted in Fig.2.

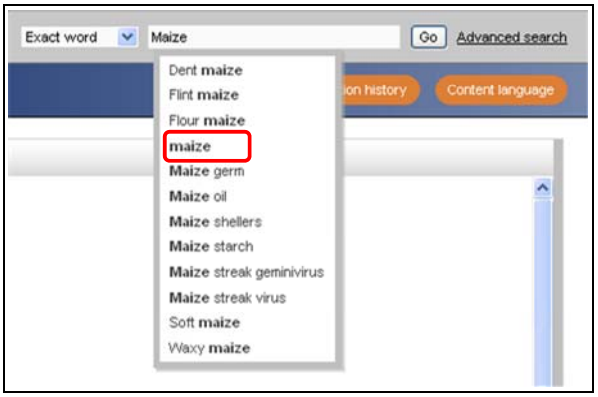

Fig 2. Searching for preferred English term "Maize"

The next step involves addition of the Malay term "Jagung" as the preferred term as illustrated in Fig. 3.

When the term is submitted by the user, it is then the publisher's role to validate and publish the proposed term as he/she seems fit. Once it is published, the newly approved term will appear as the preferred term as depicted in Fig.4.

Fig. 3. Adding New Preferred Malay Term

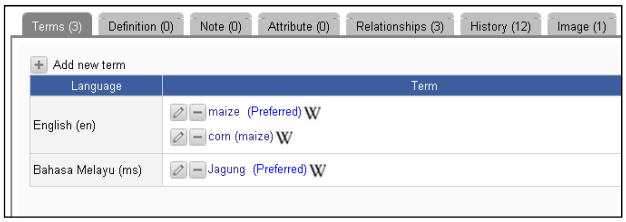

Fig. 4. Published Preferred Malay Term

2.3. Telegu Terms Translation

The main aim of translating AGROVOC into Telugu language is to develop a Telugu Agriculture thesaurus which is the first of its kind. This thesaurus would be of great help in many fields of agriculture locally used by extension personnel (development of Telugu content, brochures, pamphlets, etc. for farmers).

Initially all English AGROVOC terms were collected from the database in alphabetical order. In the alphabetical list, all available top terms were selected. Starting from these top terms, all descriptors in the hierarchy were initially translated as these are the ones which are used for indexing. Each respective translated term was checked in many authenticated books or brochures, published by Agricultural Universities, State Agricultural Department, to avoid long validation process as these resources were already validated. In case of new translations, the terms were pooled according to the subject (i.e. Agronomy, Breeding, Entomology, etc.) and sent to the concerned eminent scientists in various state departments for the validation.

After the validation of descriptors, non-descriptors were translated in the same process as mentioned above. Many of the geographical, chemical terms were simply transliterated as they represent the same concept in native language. In case of taxonomic terms, all the descriptors were transliterated, whereas the common names were translated according to their preferences in the native language.

Google transliteration [9] and Microsoft Indic Language Input Tool [10] are the tools used in assisting

with the Unicode Telegu terms translation. These Unicode Telegu terms are also tested in VocBench for their compatibility with the tool.

2.4. Thai Terms Translation

Thai AGROVOC [11], the only Thai agricultural thesaurus, is developed to enhance the retrieval efficiency and serve as a resource base for indexing and translating Thai agricultural information.

The development of Thai AGROVOC was divided into two phases. Phase 1 (2001-2005) focused on translating AGROVOC to Thai language and Phase 2 (2006-present) manages the maintenance and addition of related words in its local language. Developing the Thai AGROVOC involved the following resources:

- 37 experts in 31 fields of study with 12 lecturers and researchers working as translators
- FAO AGROVOC thesaurus
- Thai dictionary
- Special fields dictionary related to agriculture
- Text books and document on agriculture
- Vocabulary and index in articles and research paper for the past 20 years

2.4.1 AGROVOC Translation in Thai Language (Phase 1)

In this section, we detailed out the processes involved in translating the AGROVOC to Thai Language and these processes are further illustrated in Fig 5.

1. Build vocabulary collection using the vocabulary database from AGROVOC (Agrovoc.mdb) version 9 as a prototype. The vocabulary in the database is displayed in three languages (English is used for the prototype, totalling to 28,577 terms). The terms count in AGROVOC database version 9 is listed in Table 1.

Table 1. Terms Count in AGROVOC Database

Version 9

English descriptor	Non-descriptors	Deleted terms	Scope note
16,607	10,706	164	1,264

2. Categorize the 27,313 terms (vocabulary, descriptors and non descriptors, except scope note) into 47 groups in 31 subjects, based on AGRIS/CARIS subject categorization schemes and the field of expert's specialization.
3. Create terms list for specialists to translate from English into Thai. Verified them with a technical dictionary.
4. Send the translated vocabularies to respective experts for confirmation, editing and modifying if required. This step also includes other translation that the subject specialists could not do. The experts could also recommend other descriptors or synonyms.
5. Collect the translated and confirm the vocabulary from the experts.
6. Add more vocabulary by selecting terms from titles and indexes from research papers and articles in the Thai agricultural database (dating back 20 years and extracting approximately 92,605 words). Extraction for individual word was done manually due to the lack of accurate word-extraction tool in handling Thai language, since this language has no space between the words. Take for example: "การศึกษาวิธีการปลูกมะม่วงน้ำดอกไม้". It should be extracted to การศึกษา วิธี การปลูก มะม่วงน้ำดอกไม้ but it can easily be misinterpreted as การศึกษา วิธี การปลูก มะม่วง น้ำ ดอกไม้, which does not provide the same meaning.
7. Vocabulary from the previous step was regarded as a natural-language term. They were then ranked and checked for frequency based on words utilization. These words are then split into two groups: synonym to AGROVOC terms and the other as a new term.
8. The suitable term is selected as a descriptor. The rest are kept as non-descriptors (synonym). The principle criterion is to use words that are defined and recognized by the Royal Institute. Vocabulary from a technical dictionary and text was a secondary criterion. The third is the frequency of use in the literature.
9. Add the Thai vocabulary (descriptor) into the Thai AGROVOC database using the same identification as terms in English by adding a new field for Thai terms. Synonym terms are recorded separately in another table, to be used later as a non-descriptor.
10. Check for word redundancy, relationships and corrected errors.
11. Submit the vocabulary to experts for re-verification and re-editing.
12. Modify them according to the experts' recommendation and corrections.
13. Process a preliminary Thai AGROVOC.
14. Install the system to provided service and obtain feedback via the Internet.

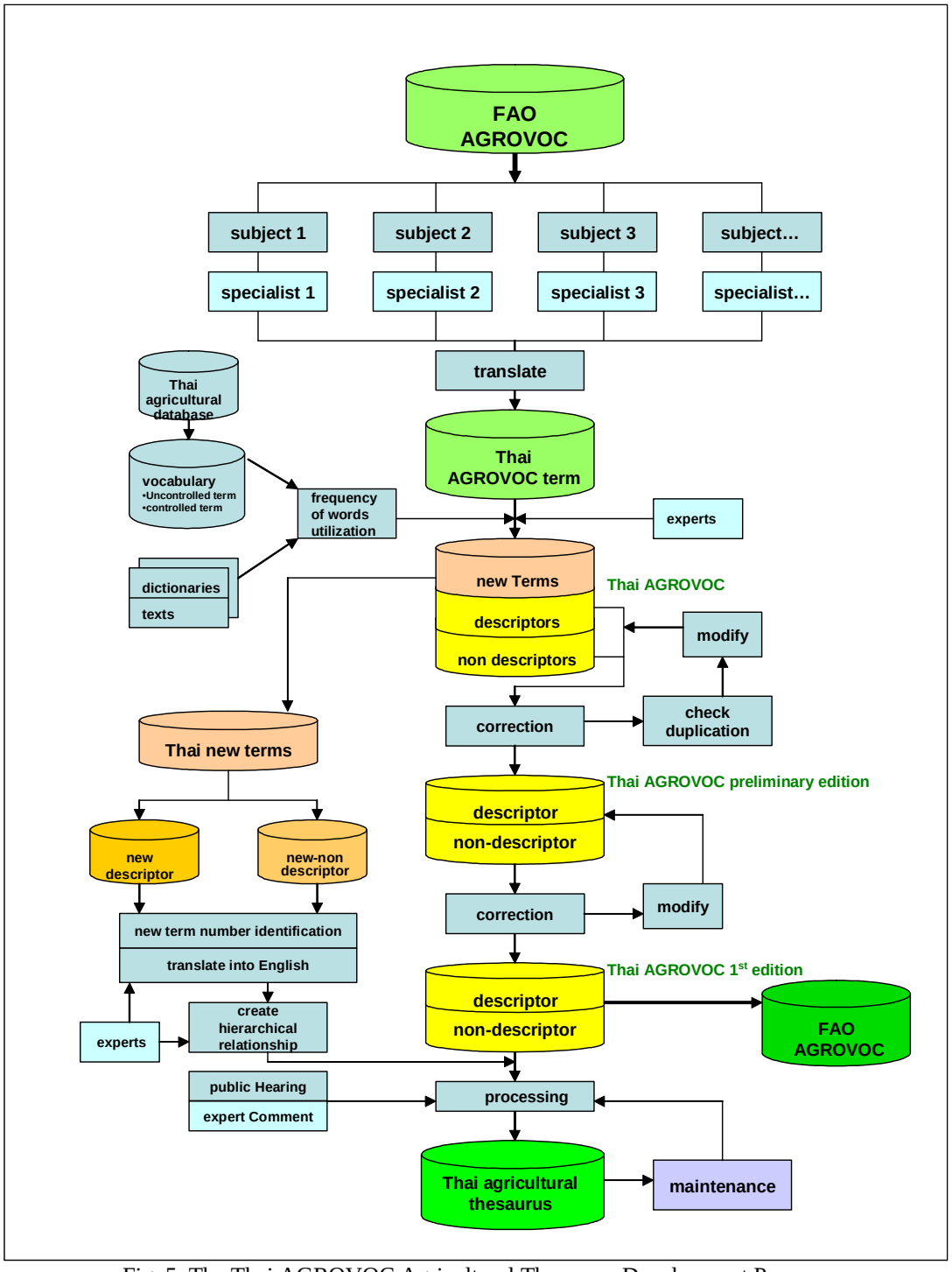

Fig. 5. The Thai AGROVOC Agricultural Thesaurus Development Process

15. Evaluate the system, verify terms and their relationship.
16. Edit data and process Thai AGROVOC first edition.
17. Install a public feedback system with Internet access.
18. Launch the service, promote it through public relations campaigns and disseminate it to interested groups of people for comments.

19. Submit the Thai AGROVOC to FAO.

2.4.2 Enhancing local terms (Phase 2)

1. Collect comments and opinions from the public; summarize and edit data.
2. Assemble proposed terms obtained from the public; as in step 7 of Phase 1, accepted vocabulary will then be taken either as new words or as synonyms.
3. Confirm those new words with experts, create a relationship with existing words and then translate new words from Thai to English.
4. Add terms with relation into Thai AGROVOC database by creating new term identification.
5. Record vocabularies, check for redundant terms and correct the errors.
6. Submit the Thai vocabulary to the Royal Institute for them to promulgate it as an approved Thai agriculture vocabulary.

Existing words are regularly modified and new words are continuously being added for maintenance. Adding new words are possible when there is a suggestion from users and by developing thesaurus maintenance tools that automatically add words and create a relationship.

2.4.3 Difficulties in Thai AGROVOC Project

There were several difficulties which were encountered during the Thai AGROVOC development project. They are as follow:

- Incompatibility between local information and vocabulary. Many descriptors in the AGROVOC are not available in the local scope of knowledge or are incomprehensible to local people. On the other hand, there are also countless local vocabularies, especially those names of plants, animals or other local being, that are available in AGROVOC.
- Difficulty in defining new terms in Thai Language. Many terms in the AGROVOC have never appeared in Thai so there is a necessity to define a new one. This endeavor is really challenging because it requires expertise and mutual collaboration from experts in that very specific field before is officially accepted and registered as new terms.
- Inconsistency in language structure. In Thai, there are no singular or plural nouns. Therefore the meaning of word with “s” and word without “s” cannot be displayed differently as in English terms in AGROVOC. For example, the terms “SEED” and “SEEDS” have different meaning in AGROVOC
- Incompatibility of meaning in synonymous words. Some word possesses more than one meaning in English but means only one thing in Thai. For example, “corn” and “maize” in Thai mean the same thing and have only one Thai term. On the other hand there are different words in Thai that means only one word in English, such as "ควาย" and "กระบือ" which in English refers to “buffaloes” only.
- The endeavor to engage many experts in various fields in addition to a lot of personnel for word extraction and data processing is a very time consuming process.
- Adding words and changing relationships are extremely difficult due to the lack of efficient tools.
- Some original terms in AGROVOC (as prototype) are not up to date; also, their relationships were sometimes incomplete, unclear or inadequate.

As such the development of Thai AGROVOC is very critical to support the local community in accessing the Agriculture knowledge base effectively. This project is financially supported by Kasetsart University’s Research and Development Institute with the technical support from FAO, which facilitated the project.

3. Challenges in Multilinguality Concept Scheme

The multilingual concept scheme focuses on building models of common knowledge which are interoperable between different languages [12, 13], so that, for example a Chinese speaker might contribute additional knowledge to a model that a German speakers started. There are many current approaches which spotlight on a universal syntax (such as OWL, SKOS etc.) so that multilinguality can be used in the concept labels. The traditional AGROVOC thesaurus has been converted into the Concept Scheme. The concept scheme has a knowledge base of around 40,000 concepts with 579523 terms organized in ontological relationships (hierarchical, associative, equivalence). These concepts were obtained through remodeling of the traditional AGROVOC thesaurus.

The AGROVOC Concept Scheme has been expressed in three different level of representation.

- *Concept* is the abstract meaning given to the group of the terms. For e.g. ‘maize’ in the sense of products.
- *Terms* are the language specific lexical form of that concept. For e.g. ‘maize’ in English, ‘Maíz’ in Spanish, or ‘Maïs’ in French or ‘玉米’ in Chinese or ‘MAIS’ in German .
- *Term* variants are the range of forms that can occur for each term. For e.g. ‘Organization’ or ‘Organization’.

Now, the concepts build the actual classification hierarchy and semantic structure of the ontology. These kinds of multilingual concept scheme bring lots of advantages in the different field. For, example, French library catalog system can use the French Concepts labels in order to index the books or other materials. Furthermore, different concept scheme have been mapped to the AGROVOC. These mapping activities are not only done in English-to-English labels but also with other languages. For example, the French subject heading RAMEAU only has concepts in French language. It has been mapped with the AGROVOC concept in French labels. These kinds of advantages enable the multilingual concept scheme to join the different resources and publish them in the Linked Open Data (LOD).

However, the multilingual concept scheme also bears some changes. Suppose there is a new user who adds terms or concepts in his/her language into the scheme, if the added terms do not have any translations in other languages, users cannot find or use these terms. To solve the issues, the scheme needs to be updated simultaneously. But, it is very difficult to keep the information up-to-date. *For example, the English term "Agribusiness" does not have any French translation. On the other hand the French term "'Haricot asperge" does not have any English translation.*

All the terms or concepts have been added to the AGROVOC by the domain experts [13, 14] in different languages, in addition to their own responsibilities and duties within their organizations. These human laborious works are very time consuming and expensive. Under unforeseen circumstances, sometimes these partner organizations are unable to update the terms frequently. For example, French terms or German have not been updated for eight years. Meanwhile, other languages are continuously being updated and expanded. Furthermore, with lack of the authenticate sources; it creates different meaning of the terms.

4. Conclusions and Enhancements to Multilinguality in AGROVOC Concept Scheme

We have presented a model of current system for managing the thesauri and concept scheme, thanks to its emphasis on collaborative workflow and its adherence to W3C standards, which is unique among open-source tools for thesauri development. In addition, we have also discussed on the importance of the historical information management systems and the new semantically-aware systems in the current maintenance process.

Furthermore, we have presented the real editorial experiences for translating the AGROVOC in different languages. It supports the new users in adding their vocabularies with the AGROVOC. According to the editors, one of the biggest challenges is to get the right definitions from the sources.

In consideration of the rising importance of linked data, the VocBench is continuously being enhanced so that it can natively support RDF/SKOS. This will have several advantages: a single triple store can then be used to both edit and disseminate linked data, removing the needs for tedious conversions. Secondly, the tool will be very useful to any community organizing their data in SKOS.

5. References

1. AGROVOC (2011). Food and Agriculture Organization of the United Nations (FAO), Rome, Italy. URL:http://aims.fao.org/standards/agrovoc (last visited September 2011).
2. Soergel, D., Lauser, B., Liang, A., Fisseha, F., Keizer, J., Katz, S.: Reengineering Thesauri for New Applications: The AGROVOC Example. Journal of Digital Information - JODI 4 (2004).
3. Gennari, J., Musen, M., Fergerson, R., Grosso, W., Crubézy, M., Eriksson, H., Noy, N., Tu, S.: The evolution of Protégé-2000: An environment for knowledge-based systems development. International Journal of Human-Computer Studies 58(1), 89–123 (2003) Protege.
4. Nublauch, H., Fergerson, R., Friedman Noy, N., Musen, M.: The Protégé OWL Plugin: An Open Development Environment for Semantic Web Applications. In : Third International Semantic Web Conference - ISWC 2004, Hiroshima, Japan (2004).
5. W3C: SKOS Simple Knowledge Organization System eXtension for Labels (SKOS-XL). In: World Wide Web Consortium (W3C). (Accessed August 18, 2009) Available at: http://www.w3.org/TR/skos-reference/skos-xl.html.
6. AGROVOC (2011). Multilingual Agricultural Thesaurus. URL: ftp://ftp.fao.org/gi/gil/gilws/aims/references/flyers/agrovoc_en.pdf (last visited September 2011).
7. Pusat Rujukan Persuratan Melayu @ DBP Malaysia (2011). URL: http://prpm.dbp.gov.my (last visited September 2011).
8. Mykamus Mobile (2011). URL: http://mykamus.com (last visited September 2011).
9. Google transliteration (2011). URL: www.google.com/transliterate (last visited September 2011).

10. Microsoft Indic Language Input Tool (2011). URL: http://specials.msn.co.in/ilit/Hindi.aspx (last visited September 2011).
11. Thai AGROVOC (2011). URL: http://pikul.lib.ku.ac.th/ (last visited September 2011).
12. Almeida, J. and Simes, A. (2006). T2O: Recycling Thesauri into a Multilingual Ontology. Fifth international conference on Language Resources and Evaluation, LREC 2006, Genova, Italy, May.
13. Gregor Thurmair: Multilingual Content Processing. In: Proceedings of the Fourth International Conference on Language Resources and Evaluation, LREC 2004, Lisbon – Portugal, 26 May - 28 May 2004, Paris: ELRA - European Language Resources Association 2004, Vol. V, XI-XVI.
14. Wikipedia (2011), Malay Language, URL: http://en.wikipedia.org/wiki/Malay_language (last visited September 2011)