Is ISO 639 enough for a multilingual thesaurus?

The AGROVOC case

Caterina Caracciolo, Gudrun Johannsen, Lavanya Kiran

Food and Agriculture Organization of the United Nations (FAO of the UN)
v.le Terme di Caracalla 1
00154 Rome
{caterina.caracciolo, gudrun.johannsen}@fao.org; lavanyakirann@gmail.com

Abstract.
In this paper we report on our analysis of multilinguality support for thesauri, with special reference to the AGROVOC thesaurus. We highlight that AGROVOC makes exclusive use of ISO 639 to tag languages, and we comment of its appropriateness for a resource such as AGROVOC. We provide examples taken from the subject area of AGROVOC to show the many cases of loan-words that appear when talking about plants and foods in different geographical areas even within countries speaking the same language. Our work is widely based on the Spanish and Portuguese cases.

Keywords: thesauri, vocabularies, multilinguality, AGROVOC, SKOS, SKOS-XL

1 Introduction

AGROVOC[1] is the corporate vocabulary of the Food and Agricultural Organization (FAO) of the United Nations and, as many similar initiatives, was born in the 1980s. The original structure of AGROVOC was a typical thesaurus structure: the notion of term is central, and well-known thesaurus relations are used to organize terms for ease of indexing and retrieval of information. First of all, the two relations Use (USE) and Used For (UF) are meant to identify the correspondences between the many terms that people may use to talk about a concept, and the single term to be used for indexing purposes. Then, the two relations Broader Term (BT) and Narrower Term (NT) are used to express hierarchies of terms, and finally, the relation Related Term (RT) is

[1] http://www.fao.org/agrovoc

used to express any type of non-hierarchical relation between terms (i.e., a generic "relatedness").

This term-centered view of thesauri stayed through the early 2000s, as the move to relational databases for storing thesauri did not challenge it. It is the move to internet-based environments, and in particular the adoption of the RDF [1] data model and SKOS vocabulary [2] for RDF2, that changed the perspective on thesauri. Now the central notion is no longer the notion of term, but rather the notion of concept, taken as the abstract meanings of terms.

SKOS introduces the generic notion of "concept" ("an idea or notion, a unit of thought. However, what constitutes a unit of thought is subjective, and this definition is meant to be suggestive, rather than restrictive. " [2]), expressed by the predicate *skos:Concept*. Following the SKOS recommendation, AGROVOC uses *skos:Concept* to indicate the "concept" behind a group of words in various languages, all then considered to be translation of one another. Following standard practices for web publishing, in particular for Linked Data publishing [4], concepts are given dereferenceable URIs (or, more plainly, URLs) so as to have a correspondence in the web environment. In case of AGROVOC, the abstract, language-independent nature of concepts is emphasized by URIs that do not refer to any specific language to name a concept. Consider for example the URI for the concept corresponding to "maize": http://aims.fao.org/aos/agrovoc/c_12332. In this view, terms are then taken as "names" or rather "labels" of a concept. The concept with URI http://aims.fao.org/aos/agrovoc/c_12332 has labels "maize", "maïs", "玉米", "ข้าวโพด" in English, French, Chinese, Hindi respectively. SKOS provides the following predicates for expressing labels: *skos:prefLabel* and *skos:altLabel*, where the former is used for descriptors (preferred terms), the latter for non-descriptors (non-preferred terms). Once terms are grouped together as labels of the same concept, the hierarchical relations BT/NT are rendered by SKOS predicates *skos:broader*, and *skos:narrower*, the domain and range of which are *skos:Concepts*. Similarly, the general non- hierarchical relation RT is expressed in SKOS by the property *skos:related*.

SKOS-XL [5], the SKOS's extension specific for the treatment of labels, is meant to provide extra support for identifying, describing and linking lexical entities. SKOS-XL allows one to attach attributes to labels (they are no longer literals, but resources just like concepts). The SKOS-XL predicates for this are then *skosxl:prefLabel* and *skosxl:altLabel*.

---

[2] Also the early adoption of OWL for managing thesauri should be mentioned, but that it is outside the scope of this paper. The interested reader may refer to [3].

2 Rendering of multilinguality in electronic vocabularies

SKOS predicates for labels, *skos:prefLabel* and *skos:altLabels*, take plain literals as values (a UNICODE string in Normal Form C) and an optional language tag expressed by the XML attribute *xml:lang*. The values of such an attribute are language identifiers as defined by the "Best current practice" on Tags for the Identification of Languages by the Internet Engineering Task Force (IETF) [6]. This is a "current practice" in that discussion on language tag within IETF is continuously ongoing.[3] Such an IEFT best practice takes as a basis the established by the International Organization for Standardization (ISO) for languages (declared in ISO 639), and adds subtags to them. Subtags can be taken from the ISO sets of codes for countries (ISO 3166), or for scripts (ISO 15924), or from both of them. For example, "tr-CY" is the IETF code for Turkish from Cyprus, while "zh-Hant-HK" is the code for Chinese written in traditional Chinese script, in Hong Kong.

AGROVOC only uses ISO 639 to specify the attribute *xml:lang*. Multilinguality is then supported because one is able to express that a term is an English term (ISO code = "en"), as opposed to, for example, a German term (ISO code = "de"). All terms used to express the same concept are kept together by being labels of the same concepts, while individual language versions of AGROVOC can always be extracted by filtering on the *xml:lang* attribute.

ISO 639 is the set of ISO codes dedicated to identifying identifying languages. ISO 639-1, the version of 2-digit long codes, is widely used in information systems, and specifically to give values to the *xml:lang* attribute. ISO 639-1 provides codes for a number of languages, independently of the countries in which they are spoken and of their official status (Fig. 1). For example, ISO 639-1 distinguishes Spanish (code "es") and Basque (code "eu"), both official languages spoken in Spain: Spanish being the official language of the whole country, while Basque being the official language of the Basque Countries and Navarra. ISO 639-1 also distinguishes French (code "fr") and Breton (code "br"), both spoken within the border of France, but Breton not having any official status. In case of linguistically similar languages, ISO 639-1 assigns the same code even if the languages are spoken in different countries. This is the case of Dutch and Flemish (code "nl"), spoken in The Netherlands and Belgium. However, there is only one ISO code for English (code "en"), only one code for Spanish (code "es") and only one code for Arabic (code "ar"), although all these languages are spo-

[3] As the dedicated mailing list shows: http://eikenes.alvestrand.no/pipermail/ietf-languages/

ken in different countries with lexical, syntactical, semantic and phonetic differences[4]. For example, the word "cilantro" is the American English word for coriander leafs, while in the US, "coriander" only refers to the seeds of the same plant.

Language	Code
Abkhazian	AB
Afar	AA
Afrikaans	AF
Albanian	SQ
Amharic	AM
Arabic	AR
Armenian	HY
Assamese	AS
Aymara	AY
Azerbaijani	AZ
Bashkir	BA
Basque	EU

Fig. 1. An excerpt of ISO 639-1 codes for languages.

The 3-digit codes ISO 639-2 expands the set of languages given a code, by including a larger number of contemporary languages, and also some no longer spoken languages, groups of languages, and artificial languages. For example, ISO 639-2 includes codes for Bemba language (Bantu language mostly spoken in Northern Zambia) and for Asturian (language spoken in the Spanish region of Asturias), but also for Akkadian (an extinct Semitic language), Old French (842-ca. 1400), for generic geographical variations of languages (Northern Frisian (code "frr"), Eastern Frisian (code "frs")), for groups of languages (Caucasian languages (code "cau"), Germanic languages (code "gem")), and for artificial languages (code "art").

Despite including more languages, ISO 639-2 adopts the same perspective as ISO 639-1. Codes continue to be assigned to "languages" as wholes, independently of the

[4] [4] Up to the point that, for example, specific textbooks are available to teach American and British English, and even bilingual dictionaries American – British English are available, c.f. http://www.englisch-hilfen.de/en/words/be-ae.htm.

lexical (or phonetical, syntactical, etc) differences existing between geographically based varieties of them.[5]

Again, consider the case of English: ISO 639-2 includes one code for all varieties of English currently spoken in all Anglophone countries, but includes a specific code for "Old English (ca. 450 - 1100)"[6]. The same happens for all varieties of Spanish (i.e., Castilian) spoken in Spanish speaking countries, and Portuguese spoken in Portugal and Brazil.[7]

When the focus is on the usage of individual words, however, this perspective shows limitations. When considering the area of interest of a thesaurus such as AGROVOC, we can notice that words used to name objects of common use, especially food and plants, vary largely from one region to the other, even within the same linguistic area. These are usually loanwords, i.e. words borrowed from a donor language and incorporated into a recipient language. For example, the fruit "avocado", is in Latin America largely called "palta", a Quechua word given by the Inca to that fruit, probably after the region they know it from, the Palta region (also name of the people leaving in the region, the Paltas). In current days, the word "palta" is widely used from Peru southward, while in Mexico it is commonly called "aguacate", after the Nahatl (an Aztecan language) word "ahuácatl". Also the case of "cilantrum" mentioned above is a case of loanword ("cilantrum" is the Spanish word for "coriander") which did not replace the original English word, but only slightly changed its meaning (by restricting it to only one part of the plant). Another example is the "Amaranthus caucatus", called "amaranto" in continental Spain, but known in Latinoamerica with a variety of different names, some of which are also specific to individual cities (and surranding areas): amaranto is "coime", "coimi", "cuimi" and "millmi" in Bolivia; "ataco morado", "sangorache", "sergorache" and "hawarcha" (Quechua word) in Ecuador; in Peru it is

---

[5] [5] Here perhaps one could discuss on when two idioms should be considered as the same or different languages. However, this debatable and debated issue falls outside the scope of this work. Here we only remark that mutual understandability is actually a widely accepted criterion of separation between languages (i.e., the answer to the question: "Would two speakers of those two languages understand each other in most cases?"), but this is too vague a notion for our purposes, as it depends on which cases are taken as "most".

[6] Despite the simplifications involved in identifying a language by means of temporal boundaries.

[7] An ISO 639-3 is currently under development, widening the list of available codes even more. However, a preliminary inspection of it reveals that the point of view understandably remained the same as the standard's previous versions.

"achis", "jataco", "coyos" (specifically in the city of Cajamarca), achita (city of Ayacucho) and kiwicha (city of Cusco).

Lexical and semantic loanwords happen very often and for several reasons, the most typical ones being the inheritance of languages previously spoken in the area, often because of lack of appropriate words in the language subsequently spoken in the same area (avocado tree is originally of Central America, Mexico). Terminological differences may also be found in more abstract, or "higher level" words: words of bureaucracy, science or technique. Portuguese is a notable example of this. Table 1 shows many differences in such registries between Portuguese used in Brazil, and Portuguese used in Portugal. From Table 1 we can notice both minor differences, only of spelling nature, such as "Rocadura", and major differences, where completely different words are used to express the same concept (e.g. "avicultura").

Table 1. Some differences between Portoguese from Brazil and Portuguese from Portugal.

ENGLISH	PT/PT	PT/BR
Agricultural planning	Planeamento agrícola	Planejamento agrícola
Reproduction control	Controlo da reprodução	Controle da reprodução
Sterile insect release	Libertação de insectos estéreis	Liberação de insetos estéreis
Canopy	Coberto arbóreo ou arbustivo	Cobertura arbórea ou arbustiva
Land clearance	Rocadura	Roçadura
agricultural research	Investigação agrária	Pesquisa agrícola
Poultry farming	Criação de aves de capoeira	avicultura
bare fallow	Pousio inculto	Pousio de/com solo nu; pousio nu
Tillage	Mobilização do solo	Trabalho do solo
Farm management	Administração exploração agrícola	Gestão agrícola
Irrigation rates	Dotação de irrigação	Taxa/grau de irrigação

3 Requirements for rendering multilinguality in vocabularies

The exclusive use of ISO 639 for tagging languages (i.e., as value of the attribute *xml:lang*) is rather limiting in AGROVOC, as it does not support the possibility of provide more accurate and specific tags for single words and phrases – as it is often needed in case of names of foods and plants. [8]

In the following, we phrase the requirements (previous requirements were given in [9]) we consider appropriate for a resource such as AGROVOC:

1. Allow for the possibility of unambiguously specifying the geographic area where a given word is used.

2. The specification of the area of use of a given word should be optional.

3. The specification of the area of use of a given word should be flexible in terms of the type of area allowed.

    - Countries, groups of countries, geographical or administrative regions should be equally available for specification.

These considerations are made here in order to guide the improvement of both AGROVOC rendering as SKOS resource, and of VocBench[9], the web-based tool used for its maintenance.

4 Conclusions

In this paper we presented a brief overview of how multilinguality is supported in AGROVOC, and highlighted some limitations we have found, related to the exclusive use of ISO 639 to tag words. We pointed out that a more fine-grained language specification should be allowed, namely one that allows for the specification of the area of use of a given term. We believe that such an extension would not only overcome limitations, but boost AGROVOC's potentialities. In fact, considering the large abundance of plant names in AGROVOC, could allow AGROVOC to become the reference resource for common and local names of plants, which is a type of information currently sparsely collected and shallowly characterized.

---

[8] Some two thirds of AGROVOC are about plant names.

[9] http://aims.fao.org/tools/vocbench-2

Future work will concentrate on further collecting and analyzing use cases for the extension of multilinguality support in AGROVOC, with more contributions expected from Portuguese and Spanish speakers from Latin America, and then moving on to addressing the issue.

References

1. Resource Description Framework (RDF) http://www.w3.org/RDF/
2. Simple Knowledge Organization System (SKOS) http://www.w3.org/2004/02/skos/
3. Caterina Caracciolo, Ahsan Morshed, Armando Stellato, Gudrun Johannsen, Johannes Keizer. Thesaurus Maintenance, Alignment and Publication as Linked Data. The AGROVOC use case. MTSR 2011.
4. Tom Heath. Chris Bizer, Linked Data: Evolving the Web into a Global data Space. http://linkeddatabook.com/editions/1.0/
5. SKOS-XL http://www.w3.org/TR/skos-reference/skos-xl.html
6. IETF 5646. Tags for Identifying Languages. http://tools.ietf.org/html/rfc5646
7. ISO 639 http://www.loc.gov/standards/iso639-2/php/code_list.php
8. Caterina Caracciolo, Margherita Sini, Johannes Keizer. Requirements for the treatment of multilinguality in ontologies within FAO. In OWLED 2007. OWL: Experiences and Directions. Christine Golbreich, Aditya Kalyanpur, Bijan Parsia (eds). Proc. of OWLED2007, CEUR workshop proceedings. Paper available at: http://sunsite.informatik.rwth-aachen.de/Publications/CEUR-WS/Vol-258/paper46.pdf .