The AGROVOC Linked Dataset

Editors: Pascal Hitzler, Kno.e.sis Center, Wright State University, Dayton, Ohio, USA; Krzysztof Janowicz, University of California, Santa Barbara, California, USA
Solicited reviews: Marta Sabou, MODUL University Vienna, Austria; Willem Robert van Hage, VU Amsterdam, The Netherlands; one anonymous reviewer

Caterina Caracciolo[a,] Armando Stellato[b], Ahsan Morshed[a], Gudrun Johannsen[a], Sachit Rajbhandari[a], Yves Jaques[a] and Johannes Keizer[a *]
[a] *Food and Agriculture Organization of the United Nations (FAO of the UN) v.le Terme di Caracalla 1, 00154 Roma, Italy*
[b] *University of Rome, Tor Vergata, Via del Politecnico 1, 00133 Rome, Italy*

Abstract. Born in the early 1980's as a multilingual agricultural thesaurus, AGROVOC has steadily evolved over the last fifteen years, moving to an electronic version around the year 2000, and embracing the Semantic Web shortly thereafter. Today AGROVOC is a SKOS-XL concept scheme published as Linked Open Data, containing links (as well as backlinks) and references to many other Linked Datasets in the LOD cloud. In this paper we provide a brief historical summary of AGROVOC and detail its specification as a Linked Dataset.

Keywords: Linked Datasets, Agriculture, Data Management

[*]Corresponding author.

1. Introduction

The AGROVOC thesaurus -- its name is a portmanteau word of Agriculture with Vocabulary -- was first published at the beginning of the 1980s by the Food and Agriculture Organization of the United Nations (FAO). At its birth AGROVOC was available in three languages (English, Spanish and French), its purpose to serve as a controlled vocabulary for the indexing of publications in agricultural science and technology, including forestry, animal husbandry, aquatic sciences, fisheries, aquaculture and human nutrition. Primary users were the FAO library and the International System for Agricultural Science and Technology (AGRIS) (http://agris.fao.org), a global public domain database coordinated by FAO and containing approximately 3.5 million bibliographic records.

In the year 2000, AGROVOC abandoned paper printing and went digital, with data storage handled by a relational database. This greatly eased maintenance. However, limitations were also experienced, especially owing to the distributed community of editors which had enlarged over the years. Also, data were available to third parties only by means of database dumps, or through web services. The models and technologies developed within the Semantic Web, and the publication methodologies and best practices promoted by Linked Open Data [1] offered the possibility to overcome these limitations. AGROVOC was remodelled using OWL [2] and then SKOS (see [3] for a detailed description of the evolution of the model). With the adoption of SKOS-XL, AGROVOC finally met the modelling requirements of a multilingual and linguistically detailed thesaurus.

Today, the AGROVOC SKOS-XL concept scheme is a LOD (Linked Open Data) Dataset composed of more than 32000 concepts available in over 20 languages (five additional languages are under development), containing up to 40,000 terms in each language.

AGROVOC is still managed by FAO, and owned and maintained by an international community of experts and institutions active in the area of agriculture. AGROVOC is widely used in specialized libraries as well as digital libraries and repositories to index content. It is also used as a specialized tagging resource for knowledge and content organization by FAO and other third-party stakeholders.

This paper provides an overall description of the AGROVOC Linked Dataset and details its maintenance and publication process. As many thesaurus managers are embracing Semantic Web technologies, we believe our work is of general interest and may serve as a use case to the community.

The rest of this paper is organized as follows: section two provides more details about publication of the linked dataset; section three presents the process followed for the generation of links between AGROVOC and relevant resources such as vocabularies, glossaries and thesauri; section four summarizes and discusses the entire data flow of AGROVOC, from maintenance to LOD publication; section five provides additional information on reported use of the AGROVOC linked Dataset and section six concludes.

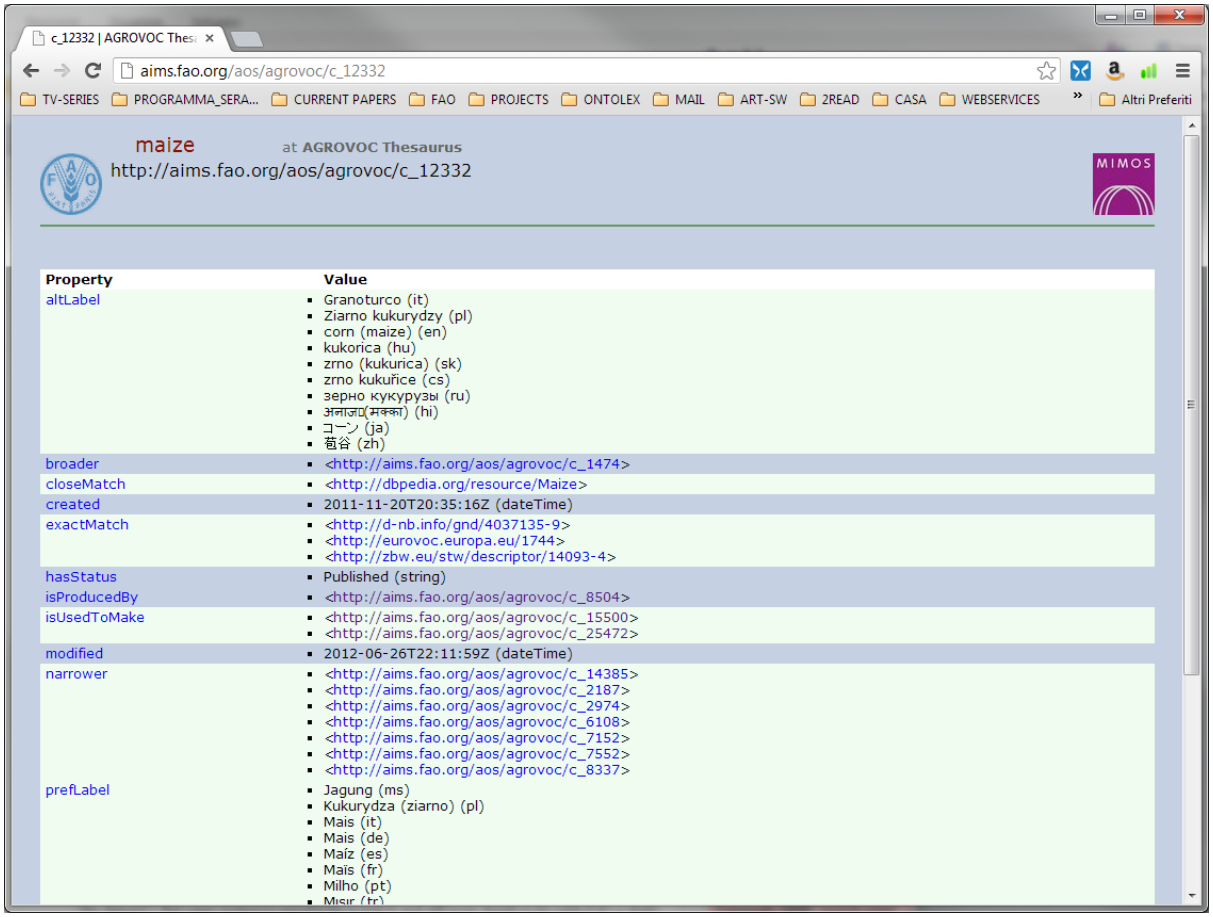

Figure 1. Part of the description of concept c_12332 (“corn”), visualized as an HTML page through Pubby

2. The Dataset

Information about AGROVOC thesaurus is available from the FAO website, at the address: http://www.fao.org/agrovoc.

The RDF version of AGROVOC has been made available as a Linked Open Dataset at the address:

http://aims.fao.org/aos/agrovoc/

and it is also available through a data dump on the project site[1] of its main editing platform, VocBench (see [4] for a detailed description of this collaborative editing tool developed by FAO and other partners). AGROVOC data is freely usable under the terms of the Creative Commons 3.0 license[2].

Content negotiation for the LOD Dataset is properly managed by the server, and clients requesting HTML content (e.g. ordinary web browsers) are returned with an HTML representation of RDF data describing the requested concept, provided by the Pubby[3] application (see Figure 1).

A description file following the VoID (Vocabulary of Interlinked Datasets) specifications [5] is available alongside the AGROVOC Linked Open Dataset: http://aims.fao.org/aos/agrovoc/void.ttl. Such a VoID file contains statistical information about the linked dataset, as well as coordinates for automatically accessing and properly querying it.

After the evolutions which its modelling exigencies dictated along the past years, AGROVOC finally found in the SKOS-XL model its perfectly-fitting dress. The SKOS-XL model, features “reified” labels which can thus be enriched with properties of their

[1] http://code.google.com/p/agrovoc-cs-workbench/downloads/list

[2] http://creativecommons.org/licenses/by/3.0/

[3] http://www4.wiwiss.fu-berlin.de/pubby/

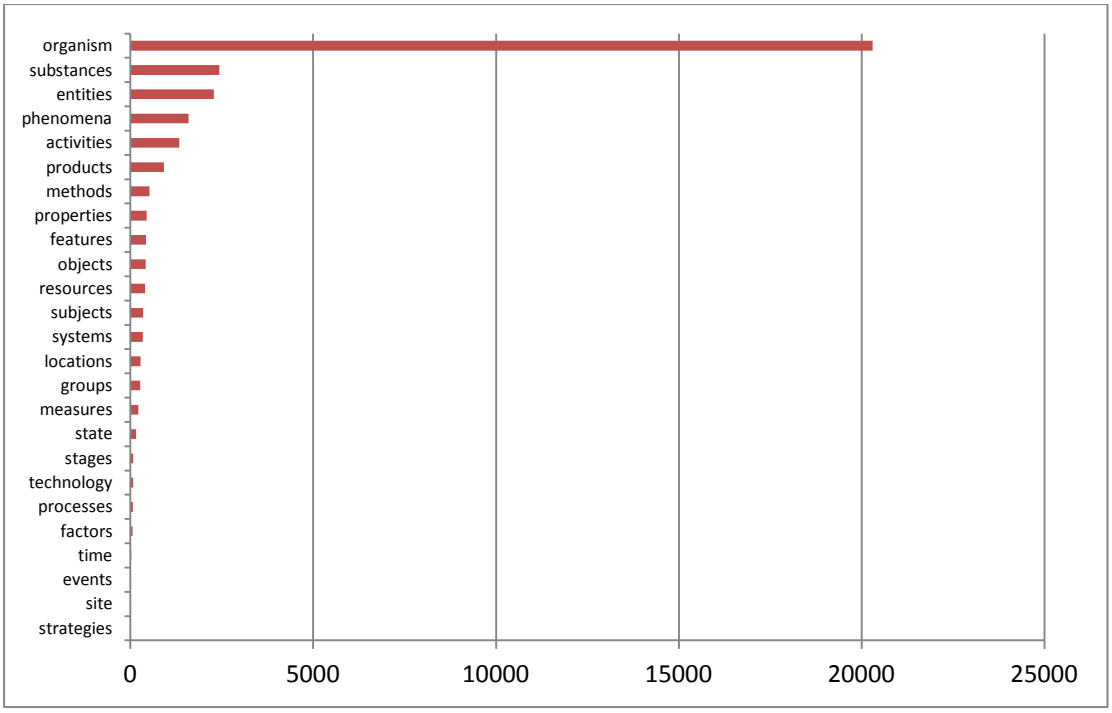

Figure 2. Distribution of the 32000+ AGROVOC concepts under the 25 top concepts of its main scheme

own. As an example, consider the concept shown in Figure 1, "maize" in English. Its Chinese preferred label[4], "玉米", is expressed in skos-xl by means of two triples[5]:

agv:c_12332 skos-xl:prefLabel agv:xl_zh_1299486844646

and

agv:xl_zh_1299486844646 skos-xl:literalForm "\u7389\u7C73"

However, note that when providing a human readable representation of the data (Figure 1), we preferred to present our visitor the actual label, instead of the reified version of it. This is why the classical skos representation of terms is also available:

agv:c_12332 skos:prefLabel "\u7389\u7C73"@zh

Beyond enabling a finer linguistic modelling of the resource (by allowing, for instance, lexical relationships across labels without involvement of the attached concepts), SKOS-XL also makes it possible to refine the grain of editorial notes at language level by adding separate information on the revision of concepts as well as of each label in each language. Thus we are able to state that "玉米" has a code, originating from the past pre-RDF versions of AGROVOC, equal to "12332". This is expressed by the triple[6]:

agv:xl_zh_1299486844646 agront:hasCodeAgrovoc "12332"

As a further example, we are also able to state that that term was created on December 12, 2002:

agv:xl_zh_1299486844646 dct:created 2002-12-12T00:00:00Z"

The Agrontology[7] seen in the above example is a compendium to AGROVOC, providing domain-specific properties for enriching the description of concepts. Agrontology is enriched with VOAF[8] (Vocabulary of a Friend) descriptors, mostly for linking it to AGROVOC (and to other Datasets adopting it, such as the FAO Biotech Glossary[9]), and to have it mentioned in the LOV Dataset. Currently, AGROVOC is undergoing a deep analysis in order to make very explicit the modelling style adopted in its various topic areas (see Figure 2). A parallel analysis of the agrontology is ongoing, with the purpose of

[4] In the classical thesaurus terminology, a preferred label corresponds to a "descriptor".

[5] asgv is used here as a prefix for the agrovoc namespace: http://aims.fao.org/aos/agrovoc/

[6] Agront is the prefix for the Agrontology's namespace, described in the next paragraph. However this predicate will be soon deprecated and replaced by the standard skos:notation one, followed by the definition of a dedicated datatype for Agrovoc codes.

[7] aims.fao.org/aos/agrontology

[8] http://purl.org/vocommons/voaf

[9] http://www.fao.org/biotech

reaching a full harmony between domain modelling and the vocabulary currently used for it.

All 32,000+ concepts of the AGROVOC thesaurus are hierarchically organized under its 25 top concepts. AGROVOC top concepts are very general and high level concepts, including concepts such as “activities”, “organisms”, “locations”, “products”, “organism”, etc. The fact that 20,000+ concepts fall under the top concept “organism”, confirms how AGROVOC is largely oriented towards the agriculture sector (see Figure 2 for a complete statistic distribution of the dataset concepts under its 25 top concepts). Other important areas of AGROVOC include “substances”, “entities”, “products” and “locations”. Beyond being listed in the AGROVOC website, the list of top concepts can be found in the VoID file for AGROVOC Linked Open Dataset. Moreover, a deep study on the current coverage of AGROVOC is under study, with the purpose of supporting human and machine users alike in their quest for information within the thesaurus and its links.

3. Linking AGROVOC to other resources

AGROVOC is today published as an Open Linked Dataset with links to thirteen vocabularies, thesauri and ontologies. Five of the linked resources are general in scope: the Library of Congress Subject Headings (LCSH) [10], , RAMEAU Répertoire d'autorité-matière encyclopedique et alphabetique unifie, Eurovoc [11], DBpedia [12], and an experimental Linked Data version of the Dewey Decimal Classification [13]. The remaining eight resources are specific to various domains: NAL Thesaurus [14] for agriculture, GEMET [15] for environment, STW for Economics, TheSoz [16] for social science and both GeoNames and the FAO Geopolitical Ontology cover countries and political regions. ASFA [17] covers aquatic science and the aptly named Biotechnology glossary covers biotechnology. These linked resources are mostly available as RDF/SKOS resources.

Table 1. Resources linked to AGROVOC.

Vocabulary	Coverage	Lang used for link discovery	#matches
EUROVOC	General	EN	1,297
DDC	General	EN	409
LCSH	General (cut on Agri.)	EN	1,093
NALT	Agriculture	EN	13,390
RAMEAU	General (cut on Agri.)	FR	686
DBpedia	General	EN	1,099
TheSoz	Social science	EN	846
STW	Economy	EN	1,136
FAO Geopol. Ontology	Geopolitical	EN	253
GEMET	Environment	EN	1,191
ASFA	Aquatic sciences	EN	1,812
Biotech	Biotechnology	EN	812
GeoNames	Gazeteer	EN	212

The linked resources were considered in their entirety barring RAMEAU, for which only agriculture related concepts were considered (amounting to some 10% of its 150 000 concepts). Candidate mappings were found by applying string similarity matching algorithms to pairs of preferred labels [6] and by using the Ontology Alignment API [7] for managing the produced matches. The common analysis language used was English in all cases except the AGROVOC - RAMEAU alignment for which French was used. Table 1 shows, for each resource linked to AGROVOC (column 1), its area of coverage (column 2), the language considered for mapping with AGROVOC (column 3), and the number of matches resulting from the evaluation (column 4, see below).

Candidate links were presented to a domain expert for evaluation in the form of a spreadsheet. Once validated, the mappings were loaded in the same triple store where the linked data version of AGROVOC is stored. All resulting validated candidate matches were considered to be skos:exactMatch, as in agv:c_12332 skos:exactMatch http://eurovoc.europa.eu/1744 ("maize" in English). Exact match has the advantage of a looser notion of equivalence than the more formal equivalence found in owl equivalence/identity properties such as owl:sameAs.

The objective when linking AGROVOC to other resources was to provide only main anchors, privileging accuracy over recall. This is why we (mostly) rejected skos:closeMatch and relied exclusively on skos:exactMatch, found by means of string-similarity techniques as opposed to more sophisticated context-based approaches. Also, the One Sense per Domain assumption which, in analogy to the “one sense per

[10] http://id.loc.gov/authorities/subjects.html
[11] http://eurovoc.europa.eu/
[12] http://dbpedia.org/
[13] http://dewey.info/
[14] http://agclass.nal.usda.gov/
[15] http://www.eionet.europa.eu/gemet
[16] http://www.gesis.org/en/services/tools-standards/social-science-thesaurus/
[17] http://www4.fao.org/asfa

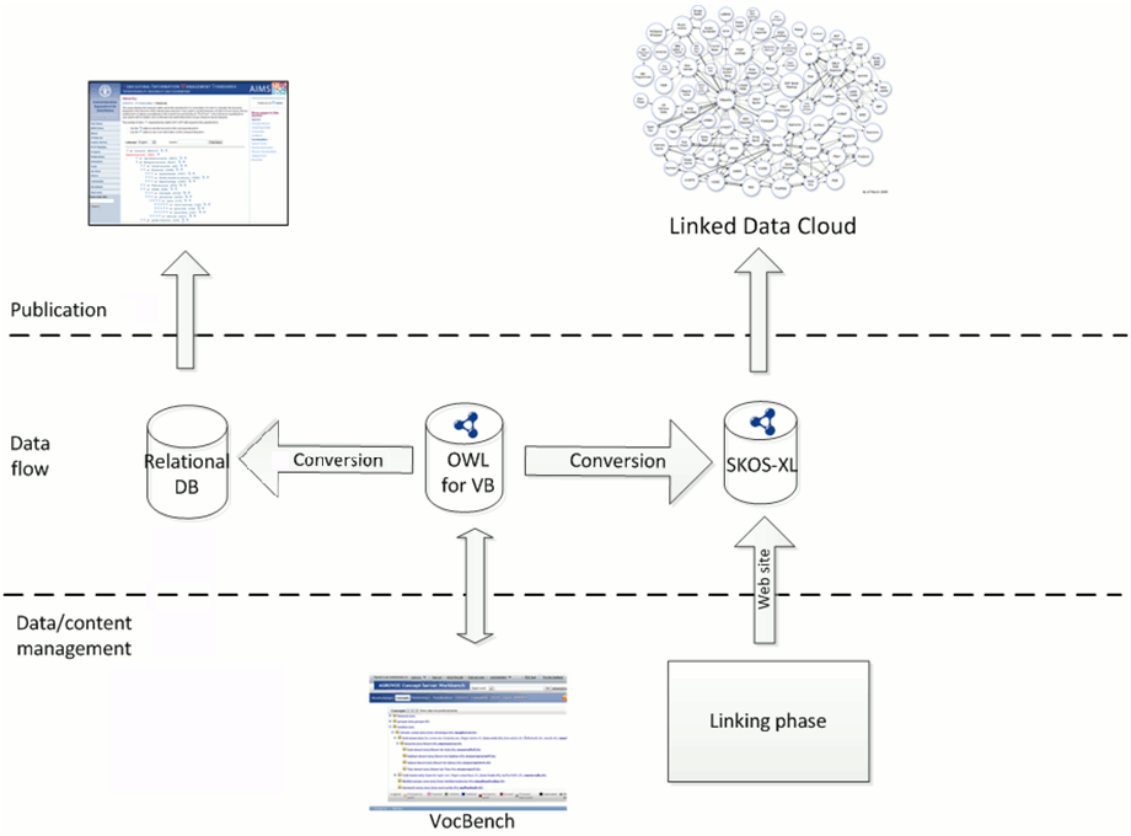

Figure 3. Overview of the process for publishing AGROVOC as linked data

discourse" [8] assumption, specifies that "the more specialized a domain is, the less is the influence of word sense ambiguity", supports our claim that (in our case) similar strings correspond to equivalent meanings. The use of more sophisticated approaches might have contributed to filtering out potential results more than widening their number (thus incrementing precision over recall), however this potential loss of precision was well compensated by the manual validation of candidate links by a domain expert.

4. AGROVOC LOD data flow

Figure 3 (next page) provides a high-level view of the entire AGROVOC's maintenance process and its publication as linked data. The figure emphasises the three levels of data maintenance (bottom layer), data storage (middle layer), and data publication (top layer).

The relational database is still necessary as many existing applications interface with this legacy model via SQL. Such conversions are thus needed to synchronize the data accessed by editors using legacy tools. This duplication of data repositories and the consequent data conversions is obviously not ideal and in principle should be limited. On the other hand, AGROVOC has supported a worldwide community of users (people and institutions) for decades who have developed a number of applications relying on the legacy relational model: these conversion steps are thus currently unavoidable and give an idea of the complexity inherent to historic, distributed collaboration scenarios. Elaborate procedures are rendered necessary, and the conversion effort, modelling issues and information needs are just the tip of the iceberg compared to the real effort spent in content and services maintenance.

Several conversion steps are then present in the AGROVOC lifecycle. Note that this data flow is not always monotonic. Although the main authoring tool

is VocBench, contributions to AGROVOC may also occasionally come from legacy formats such as spreadsheets and SQL files. This updated content is thus contributed separately (through different modalities) and then merged to produce a new copy.

When a VocBench version is finalized with contributions coming from different sources and formats, it is then converted back to the relational DB for legacy applications. At the same time, a SKOS-XL version is produced and enriched with information, such as metadata descriptors from the VoID vocabulary to feed the LOD endpoint with updated data. Currently, no versioning info for the dataset as a whole (i.e. which AGROVOC release a client is accessing), is explicitly reported inside its triples, while editorial notes provide fine-grained details about its content, with creation and modification dates for all concepts and labels present in the dataset.

The linked data version of AGROVOC is now available online thanks to a collaboration between FAO and MIMOS Berhad. Data is stored in an RDF triple store (Allegrograph[18]) hosted on a server in Kuala Lumpur, Malaysia. A SPARQL endpoint, combined with http resolution of AGROVOC entities, allows for publication as linked data. The Pubby service mentioned in section 2 is also hosted by MIMOS. Both RDF and HTML access are resolved through content negotiation on FAO servers and redirected to the proper MIMOS service.

5. Usage

During the more than 30 years of its existence, AGROVOC has seen a growing community of users exploiting its content for a progressively wider set of uses. In this section we report the more important uses of which we are aware. In some cases they exploit AGROVOC to give further explicit contribution to the LOD cloud itself, while in others, the availability of AGROVOC data as LOD will foster wider access possibilities and probably see an increase in use as the number of potential users augment due to their interaction with the entirety of the LOD cloud and to AGROVOC's position within it.

5.1. Data.fao.org

In 2011, following a wave of enthusiasm caused by Linked Open Data initiatives and benefiting from the successful experiences of the AIMS group working on AGROVOC and other concept schemes and vocabularies ported to the Semantic Web, FAO's Information Technology Division (CIO) chose to add a taste of Semantic Web to their ambitious data integration project *data.fao.org.* The project which launches publicly in December 2012, brings much of FAO's statistical, textual and geographical data under one roof, fostering data integration and harmonization first within FAO itself, and later publicly via LOD. The models which are being exploited are many, mainly covering domain representation (OWL and SKOS as core modelling vocabularies), flanked by "standard" vocabularies such as FOAF [9]) and statistical data reporting (Data Cube Vocabulary [10]).

AGROVOC, the first FAO resource to embrace the Semantic Web and to be published on the LOD cloud, was chosen as a common, controlled vocabulary for tagging the information resources (documents, media etc..) in data.fao.org. AGROVOC will also act as an interlingua to easily match RDF resources from different datasets, which still maintain a certain independence and which thus expose potential overlaps with other datasets. A new, potentially wider, set of linksets on a star configuration with AGROVOC in the centre will be elaborated for establishing a global interconnected network of resources within FAO.

5.2. Agrovoc Web Services

AGROVOC and other vocabularies hosted on VocBench (e.g. Journal Authority Descriptions) have for some years been supported by an extensive set of SOAP web services[19] that allow others to seamlessly integrate vocabularies into their applications. The services support a variety of keyword searches and most methods return either SKOS or TXT.

The web services are in use by FAO's library, terminology, translation, knowledge management and capacity development groups for indexing and to aid in the translation of FAO documents. A number of CMSs as well as several FAO-supported digital repository solutions also access these services, such as AgriDrupal and AgriOcean Dspace. The publication of AGROVOC as a LOD dataset will probably make many web services users move to standard SPARQL queries, up to a point in which many of these services will be abandoned. Other services instead, due to the complex nature of the results they offer, or of the queries they implement (and to the local optimization

[18] http://www.franz.com/agraph/allegrograph

[19] http://aims.fao.org/tools/vocbench-2/web-services

which may have been brought to them) will continue to flank the standard SPARQL offer.

5.3. Other users/stakeholders

Research and academia also commonly make use of AGROVOC in their work. AGROVOC is used for indexing the library collections of most CGIAR[20] centres and numerous agricultural research institutions worldwide. Of recent note are the AGROVOC Topic Map developed in Kyoto[21], and the integration of AGROVOC into two recent indexers, HIVE[22] and MAUI. Having AGROVOC on the LOD means having a ubiquitous centralized index that is easily browsable to support the meaning of adopted index tags and to add coherence to their adoption across different uses within the same and possibly across different resources, as in the above cited examples.

6. Conclusions

The lifecycle of AGROVOC, from evolution and maintenance, to alignment with other thesauri and finally to publication as linked data is supported by an entire development chain, consisting of users engaged in a workflow supported by specialized tools. In particular, the re-modelling of AGROVOC using OWL and SKOS and its publication as linked data imply a series of discrete steps requiring a mixture of domain experts, terminologists, ontologists and software developers. These roles must in turn be supported by a series of tools: editors and workflow managers such as VocBench, triple stores and SPARQL endpoints such as Allegrograph, RDF visualizers such as Pubby, and RDF APIs such as the Ontology Alignment API. In addition, careful attention must be paid to managing the support and migration of legacy applications tied to non-RDF models.

In the current process, both historical information systems and new semantically-aware systems play a role: a streamlined sequence of conversion steps is thus impossible to realize. Support for previous versions and their user base is in fact a business process requirement that cannot be ignored. Work is ongoing to provide training to AGROVOC editors, organizing workshops for data managers, and in improving the functionalities of the VocBench environment so that it can be used by all. Also, the quality control of AGROVOC content (for both its terminological and structural aspects) is continuous.

In this light, the immediate issues to address include the improvement of off-line VocBench editing (to address the needs of low-bandwidth users), continual VocBench usability improvements (which includes adapting its user interface to various language communities), and the completion of the revision and standardization of the AGROVOC model. This final point is expected to improve the efficiency of VocBench, and to streamline editors' work.

In consideration of the rising importance of linked data, development continues on VocBench so that it may natively support RDF/SKOS. This will have several beneficial effects: a single triple store can then be used to both edit and disseminate linked data, removing the need for tedious conversions. Secondly, the tool will be of use to any community organizing their data in SKOS. Another planned development is the integration within VocBench of the cross-vocabulary alignment functionalities that are currently hosted in Eclipse. This will integrate the alignment workflow with the overall AGROVOC editing workflow.

The process followed to maintain, align and publish AGROVOC as linked data is repeatable. It is hoped that this overview can be useful to others with similar goals or problems.

Acknowledgments

The work described in this paper could have not been possible without the collaboration of a number of people. We wish to thank our colleagues Lim Ying Sean, Prashanta Shrestha, Lavanya Neelam, Jérôme Euzenat, Stefan Jensen, Antoine Isaac, Søren Roug and Thomas Baker.

References

[1] C. Bizer, T. Heath and T. Berners-Lee, "Linked Data - The Story So Far," *International Journal on Semantic Web and Information Systems (IJSWIS)*, Special Issue on Linked Data, vol. 5, no. 3, pp. 1-22, 2009.

[2] D. Soergel, B. Lauser, A. Liang, F. Fisseha, J. Keizer e S. Katz, «Reengineering Thesauri for New Applications: The AGROVOC Example,» *Journal of Digital Information - JODI*, vol. 4, 2004.

[3] C. Caracciolo, A. Stellato, S. Rajbahndari, A. Morshed, G. Johannsen, J. Keizer e Y. Jacques, «Thesaurus Maintenance, Alignment and Publication as Linked Data,» *International*

[20] Consultative Group on International Agricultural Research
[21] http://infos.net.cias.kyoto-u.ac.jp:8083/agrovoc/index.jsp
[22] http://hive.nescent.org

Journal of Metadata, Semantics and Ontologies (IJMSO), vol. 7, n. 1, pp. 65-75, Tuesday, 14 August 2012.

[4] A. Stellato, A. Morshed, G. Johannsen, Y. Jacques, C. Caracciolo, S. Rajbhandari, I. Subirats., and J. Keizer. "A Collaborative Framework for Managing and Publishing KOS," in *The 10th European Networked Knowledge Organisation Systems (NKOS) Workshop*, Berlin, Germany, 2011. Available online on: http://www.fao.org/docrep/article/am814e.pdf. Workshop information available on: https://www.comp.glam.ac.uk/pages/research/hypermedia/nkos/nkos2011/

[5] K. Alexander, R. Cyganiak, M. Hausenblas and J. Zhao, "Describing Linked Datasets with the VoID Vocabulary (W3C Interest Group Note)," World Wide Web Consortium, 3 March 2011. [Online]. Available: http://www.w3.org/TR/void/. [Accessed 16 May 2012].

[6] W. W. Cohen, P. Ravikumar and S. E. Fienberg, "A comparison of string distance metrics for name-matching tasks," in *Proceedings of the IJCAI-2003 Workshop on Information Integration on the Web (IIWeb-03)*, S. Kambhampati and C. A. Knoblock, Eds., Acapulco, Mexico, pp. 73-78, 2003. Available online at: http://www.isi.edu/info-agents/workshops/ijcai03/papers/Cohen-p.pdf

[7] J. David, J. Euzenat, F. Scharffe and C. Trojahn dos Santos, "The Alignment API 4.0," *Semantic Web Journal*, vol. 2, no. 1, pp. 3-10, 2011.

[8] W. Gale, K. Church e D. Yarowsky, «A Method for Disambiguating Word Senses in a Large Corpus,» *Computers and the Humanities*, n. 26, pp. 415-439, 1992.

[9] "Friend Of A Friend Ontology (FOAF)," [Online]. Available: http://xmlns.com/foaf/0.1/.

[10] J. Tennison, "The RDF Data Cube Vocabulary - W3C Working Draft," 5 April 2012. [Online]. Available: http://www.w3.org/TR/vocab-data-cube/. [Accessed 16 May 2012].