The AGROVOC Linked Dataset

Editor(s): Name Surname, University, Country
Solicited review(s): Name Surname, University, Country
Open review(s): Name Surname, University, Country

Caterina Caracciolo[a,] Armando Stellato[b], Ahsan Morshed[a], Gudrun Johannsen[a], Sachit Rajbahndari[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 eighties as a multilingual authority file of agricultural index terms, AGROVOC has steadily evolved these last thirty years, moving to an electronic version around the year 2000 and shortly thereafter embracing the Semantic Web. Today AGROVOC is a SKOS-XL concept scheme published as Linked Open Data cloud, 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

1. Introduction

AGROVOC (Agriculture Vocabulary) was first published in the early eighties by the Food and Agriculture Organization of the United Nations (FAO) as a multilingual (English, Spanish and French) collection of index terms to be used in cataloguing agricultural publications.

The coverage of AGROVOC included all areas of interest to FAO, e.g. agriculture, fisheries, nutrition, forestry and environment, and was first used in indexing AGRIS (International System for Agricultural Science and Technology), a global public domain database facilitated by FAO, grown now close to three million structured bibliographical records.

In the year 2000, AGROVOC abandoned paper printing and went digital, with storage handled by a relational database. This was a great improvement in terms of ease of maintenance. However, some limitations were also experienced, especially owing to the distributed community of editors which over the years had proliferated. Also, data were available to third parties only by means of database dumps, or through web services.

Since 2002, the AIMS (Agriculture Information Management Standards) group of FAO, which heads the maintenance and evolution of AGROVOC, has shown high interest in the Semantic Web and its premises (and promises) of interconnectable open datasets shared on the Web; a progressive shift to the this new approach was thus planned. 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 the limitations of AGROVOC maintenance and exploitation. In the following years (see [2] for a detailed description of the evolution of the model) AGROVOC, by first being remodelled in OWL [3] and later in SKOS, was finally able to meet the full modelling requirements of a multilingual and linguistically detailed thesaurus by using SKOS-XL.

Today, the AGROVOC SKOS-XL concept scheme is a LOD (Linked Open Data) Dataset composed of 32035 concepts available in over 20 languages (5 more languages are under development), with an av-

*Corresponding author. E-mail: editorial@iospress.nl. Check if the checkbox in menu *Tools/Options/Compatibility/Lay out footnotes like Word 6.x/95/97* is selected if you make a footnote for the corresponding author.

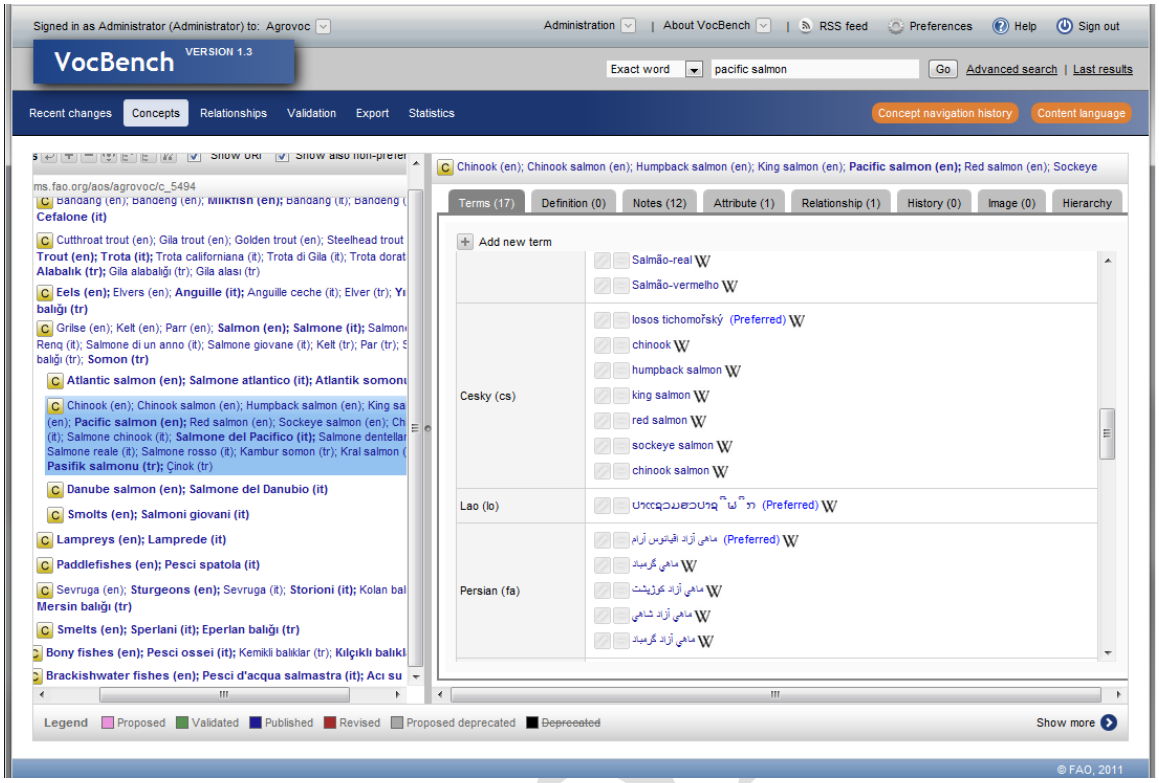

Figure 1. VocBench v1.3. User interface showing a fragment of AGROVOC.

erage of 40,000 terms in each language. AGROVOC is still managed by FAO, and owned and maintained by an international community of individual 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 as well as third party stakeholders.

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

The rest of this paper is organized as follows. Section 2 presents the editing and workflow management of AGROVOC and its maintenance tool VocBench. Section 3 presents the process followed for the generation of links between AGROVOC and relevant resources such as vocabularies, glossaries and thesauri. Section 4 summarizes and discusses the entire data flow of AGROVOC, from maintenance to LOD publication. Section 5 provides additional information on reported use of the AGROVOC linked Dataset and section 6 concludes.

2. AGROVOC Main Editing Service: VocBench

The many changes required by the passage from a closed relational database to an open data environment inspired the need for a distributed and collaborative editing environment. Special attention to user roles, strict validation procedures and diversification of editing rights (administrative, concept or modelling) was also required. A few ontology editors such as Protégé were already available [4,5] but none met AGROVOC's extensive requirements.

Work began in 2004 on the AGROVOC Concept Server Workbench, a.k.a "The WorkBench", a web-based, fully multilingual vocabulary editor supporting distributed collaboration structured into a formalized workflow. The successor of that tool is VocBench. VocBench improves on its predecessor as it fully supports a formalized editing workflow by user role and language, including a fine-grained change-tracking mechanism that allows individuals and organizations to contribute to AGROVOC while main-

taining provenance information regarding authorship. The SKOS-XL model, featuring "reified" labels which can thus be enriched with properties of their own, beyond enabling a finer linguistic modelling of the resource (by allowing, for instance, lexical relationships across labels without involvement of the attached concepts) 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.

Moreover, support to multilinguality in search, visualization and editing is fundamental to VocBench. Figure 1 presents a screenshot of the VocBench user interface showing a fragment of AGROVOC.

These features have made the interest around VocBench grow, which has in turn contributed to the refinement of VocBench requirements. VocBench is no longer an AGROVOC-only editor, and its community of users has grown beyond the one originally envisaged. Today VocBench is also used to maintain the FAO Biotechnology Glossary and much of the bibliographic metadata vocabularies used by FAO.

The use of VocBench has had a positive effect on the management of AGROVOC. It gave new impetus to the AGROVOC translations that were in progress by promoting a distributed, collaborative workflow, while today fostering a handful of other new translations and as noted above other vocabularies, a pleasing side-effect that was not part of the original vision.

The current official release of VocBench (version 1.3) still relies internally on the original OWL-based meta-model used to design AGROVOC in 2005. However, this meta-model covers all of the expressive power of the later standardized SKOS and SKOS-XL languages, thus VocBench is able to losslessly export data into SKOS/SKOS-XL. Under active development, the next release of VocBench, will be a major milestone (2.0), with native support for standard OWL, SKOS and SKOS-XL as well as broad triple-store support via the OWLART data access API from the University of Tor Vergata.

3. Linking AGROVOC to other resources

AGROVOC in the Linked Data Cloud is currently aligned with thirteen vocabularies, thesauri and ontologies in areas related to the domains it covers. Six of the linked resources are general in scope: the Library of Congress Subject Headings (LCSH), NAL Thesaurus, RAMEAU Répertoire d'autorité-matière encyclopedique et alphabetique unifie, Eurovoc, DBpedia, and an experimental Linked Data version of the Dewey Decimal Classification. The remaining seven resources are specific to various domains: GEMET on the environment, STW for Economics, TheSoz is about social science and both GeoNames and the FAO Geopolitical Ontology covers countries and political regions. ASFA covers all 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	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 thesauri 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).

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.

Our objective when linking AGROVOC to other resources was to provide only main anchors, privileg-

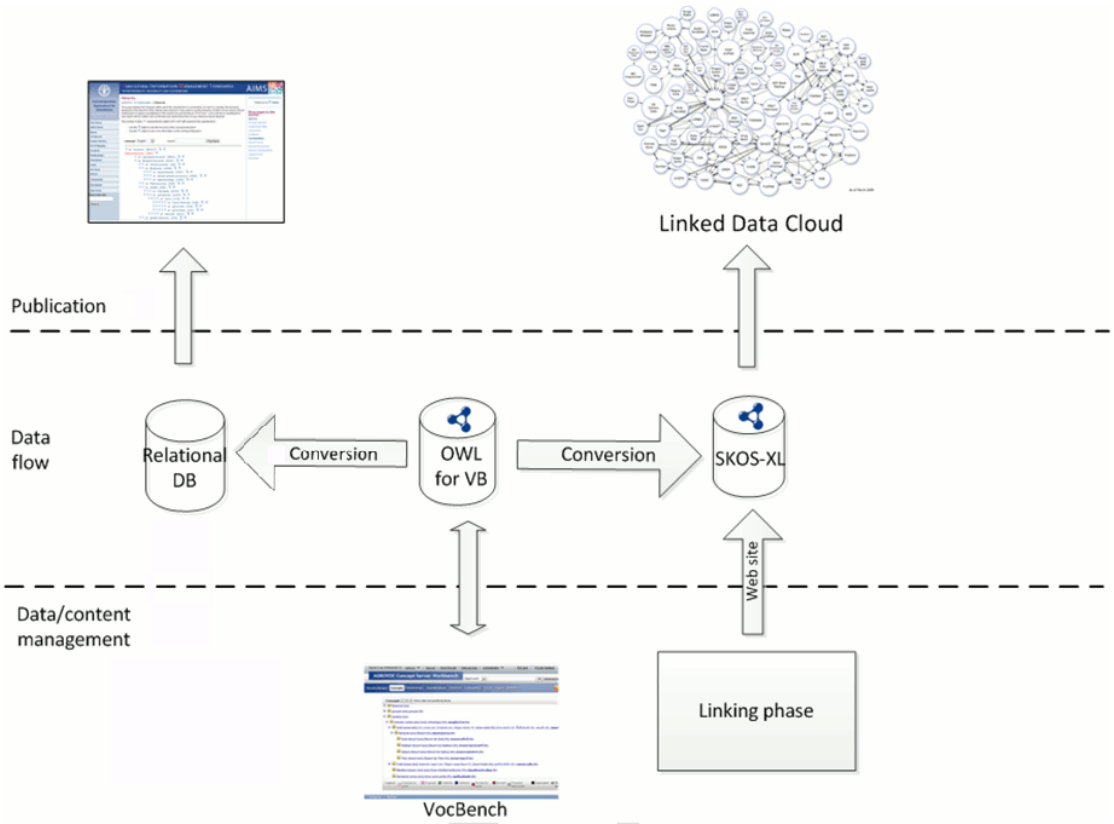

Figure 2. Overview of the process for publishing AGROVOC as linked data

ing accuracy over recall. This is why we only used exactMatch, found by means of string-similarity techniques as opposed to more sophisticated context-based approaches. Also, the One Sense per Domain hypothesis [8] 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 2 provides a high-level view of the entire AGROVOC 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 used as many existing applications interface with this legacy model. Such conversions are needed to synchronize the data accessed by editors using legacy tools that are still in operation. 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.

4.1. AGROVOC as a Linked 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[1]) hosted on a high-performance server in Kuala Lumpur, Malaysia. A SPARQL endpoint, combined with http resolution of AGROVOC entities, allows for publication as linked data. The HTML representation of linked data is made available through a customized version of Pubby (to provide more readable labels for properties, hide redundant data, etc.). The HTML representation of AGROVOC as linked data is publicly accessible[2]. Both RDF and HTML access are resolved through content negotiation on FAO servers and redirected to MIMOS. Finally, a description of AGROVOC following VoID (Vocabulary of Interlinked Datasets) specifications, detailed in [9], is also provided[3].

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.

5.1. Data.fao.org

In 2011, following a wave of enthusiasm provoked by Linked Open Data initiatives together with 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 Semantic Web approaches to their ambitious data integration project *data.fao.org*. This project that will launch publicly within 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 with typical "standard" vocabularies such as FOAF [10]) and statistical data reporting (Data Cube Vocabulary [11]).

AGROVOC, the first FAO resource to embrace the Semantic Web and publish on the LOD cloud has been 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[4] 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.

5.3. Other users/stakeholders

Research and academia also commonly make use of AGROVOC in their work. AGROVOC is used for indexing libraries of most CGIAR centres and numerous agricultural research institutions worldwide. Of recent note are the AGROVOC Topic Map developed in Kyoto[5], and the integration of AGROVOC into two recent indexers, HIVE[6] and MAUI

[1] http://www.franz.com/agraph/allegrograph
[2] http://aims.fao.org/standards/agrovoc/linked-open-data
[3] http://aims.fao.org/aos/agrovoc/void.ttl
[4] http://aims.fao.org/tools/vocbench-2/web-services
[5] http://infos.net.cias.kyoto-u.ac.jp:8083/agrovoc/index.jsp
[6] http://hive.nescent.org

6. Conclusions

The whole lifecycle of AGROVOC, ranging from its evolution and maintenance, to its alignment with other thesauri and finally to its 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 it 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 set 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 OWLART and 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] Christian Bizer, Tom Heath, and Tim 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] Caterina Caracciolo et al., "Thesaurus Maintenance, Alignment and Publication as Linked Data," *International Journal of Metadata, Semantics and Ontologies (IJMSO) [accepted for publication]*, 2012.

[3] Dagobert Soergel et al., "Reengineering Thesauri for New Applications: The AGROVOC Example," *Journal of Digital Information - JODI*, vol. 4, 2004.

[4] John Gennari et al., "The evolution of Protégé-2000: An environment for knowledge-based systems development," *International Journal of Human-Computer Studies*, vol. 58, no. 1, pp. 89–123, 2003, Protege.

[5] Holger Knublauch, Ray W. Fergerson, Natasha Friedman Noy, and Mark, A. Musen, "The Protégé OWL Plugin: An Open Development Environment for Semantic Web Applications," in *Third International Semantic Web Conference - ISWC 2004*, Hiroshima, Japan, 2004.

[6] W. W. Cohen, P. Ravikumar, and S. E. Fienberg, "A comparison of string distance metrics for name-matching tasks," in *IJCAI-2003*, 2003.

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

[8] W. Gale, K. Church, and D. Yarowsky, "A Method for Disambiguating Word Senses in a Large Corpus," *Computers and the Humanities*, no. 26, pp. 415-439, 1992.

[9] Keith Alexander, Richard Cyganiak, Michael Hausenblas, and Jun Zhao. (2011, March) World Wide Web Consortium (W3C). [Online]. http://www.w3.org/TR/void/

[10] Friend Of A Friend Ontology (FOAF). [Online]. http://xmlns.com/foaf/0.1/

[11] Jeni Tennison. (2012, April) World Wide Web Consortium (W3C). [Online]. http://www.w3.org/TR/vocab-data-cube/