
Thesaurus maintenance, alignment and publication as linked data: the AGROVOC use case

Caterina Caracciolo*

Food and Agriculture Organization of the United Nations (FAO of the UN),
v.le Terme di Caracalla 1, 00154 Rome, Italy
E-mail: caterina.caracciolo@fao.org

*Corresponding author

Armando Stellato

University of Rome, Tor Vergata,
Via del Politecnico 1, 00133 Rome, Italy
E-mail: stellato@info.uniroma2.it

Sachit Rajbahndari, Ahsan Morshed,
Gudrun Johannsen, Yves Jaques
and Johannes Keizer

Food and Agriculture Organization of the United Nations (FAO of the UN),
v.le Terme di Caracalla 1, 00154 Rome, Italy
E-mail: sachit.rajbahndari@fao.org
E-mail: ahsan.morshed@fao.org
E-mail: gudrun.johannsen@fao.org
E-mail: yves.jaques@fao.org
E-mail: johannes.keizer@fao.org

Abstract: The AGROVOC multilingual thesaurus maintained by the Food and Agriculture Organisation (FAO) of the United Nations is now published as linked data. In order to reach this goal AGROVOC was expressed in Simple Knowledge Organisation System (SKOS) and its concepts provided with dereferenceable URIs. AGROVOC is now aligned with ten other multilingual Knowledge Organisation Systems (KOS) related to agriculture, using the SKOS properties exact match and close match. Alignments were automatically produced in Eclipse using a custom-designed tool and then validated by a domain expert. The resulting data is publicly available to both humans and machines using a SPARQL endpoint together with a modified version of Pubby, a lightweight front-end tool for publishing linked data. This paper describes the process that led to the current linked data AGROVOC and discusses current and future applications and directions. This paper extends a shorter version presented at MTSR 2011.

Keywords: AGROVOC; mapping; agriculture; linked data; VocBench, thesauri alignment; OWL; Web Ontology Language; SKOS; simple knowledge organisation system.

Reference to this paper should be made as follows: Caracciolo, C., Stellato, A., Rajbahndari, S., Morshed, A., Johannsen, G., Jacques, Y. and Keizer, J. (2012) 'Thesaurus maintenance, alignment and publication as linked data: the AGROVOC use case', *Int. J. Metadata, Semantics and Ontologies*, Vol. 7, No. 1, pp.65–75.

Biographical notes: Caterina Caracciolo, PhD, is an Information Specialist at FAO since 2006. She has been involved in the development of metadata solutions for textual as well as statistical data in FAO. Her scientific background is in Information Retrieval and Knowledge Management and her main interests lay in the area of information and knowledge management. She is currently involved in the quality assurance process of AGROVOC, in the development of the AGROVOC Linked Data project and in the development of the web-based vocabulary editor VocBench.

Armando Stellato, PhD, is Research Associate at the University of Rome, Tor Vergata, where he researches and teaches in the field of Knowledge Representation and Knowledge Based Systems. He is Author of more than 50 publications and has been Member of the program committees of over 20 international scientific conferences and workshops. He is involved in a framework research agreement between his university and the European Space Agency (ESA) and is also Consultant

at the Food and Agriculture Organisation (FAO) of the United Nations as Semantic Architect, working on all aspects related to maintenance and publication of FAO RDF vocabularies such as AGROVOC, Biotech and Authority Control and of related software.

Sachit Rajbhandari, ME is an Information Management Specialist at FAO since 2007. He is Lead Developer of the web-based vocabulary-editing tool VocBench. He received his master degree in Information and Communication Technologies from Asian Institute of Technology (AIT), Thailand in 2006 and Master degree in Business Studies from Tribhuvan University, Nepal in 2003. His main research interests are in the area of information and knowledge management, semantic web, data mining and application development.

Ahsan Morshed, PhD, is an Information and Knowledge Management Specialist at FAO. He manages the publication of AGROVOC in various formats, is responsible for aligning AGROVOC with other thesauri and publish them as Linked Data. He is Author of 15 publications and Member of 4 scientific committee. He is Member of DC task group. His interest in vocabulary mapping, Linked Data, data provenance and knowledge management.

Gudrun Johansen has worked for over 20 years in the area of agricultural information and knowledge management, particularly on international metadata standards and Knowledge Organisation Systems (KOS). She is currently working as Information Systems Officer in the Office of Knowledge Exchange, Research and Extension (OEK) at FAO, responsible for the content management of the AGROVOC Concept Scheme and mapping AGROVOC to other KOS. Before joining FAO, she worked at ZADI, the Central Agency for Agricultural Documentation and Information in Bonn, Germany, as cataloguer and subject indexer for publications in the agricultural domain. As a Member of the international AGROVOC working group for updating AGROVOC, she participated in the translation of AGROVOC into German language. She holds an MSc degree in Agronomy with specialisation in Plant production and protection from Bonn University in Germany and participated in a post-graduate course in business management and computer science.

Yves Jaques, MSc SoftDev, is an Information and Knowledge Management Officer at the Food and Agriculture Organization (FAO) of the United Nations. Focusing on IT and IM/KM strategies and solutions in support of FAO's mission to end world hunger he has participated in cutting-edge knowledge engineering projects (NeON, D4Science) as well as bricks-and-mortar data management and capacity development in challenging Sub-Saharan environments (CountrySTAT). Working to build crosswalks between resources and systems he currently manages development of his branches linked data infrastructure while also representing the department in a number of cross-cutting organisational IT and IM/KM projects and initiatives.

Johannes Keizer, PhD, leads the Agricultural Information Management Standards and Services (AIMS) team in FAO. The AIMS team takes care of the multilingual AGROVOC thesaurus, facilitates the AGRIS (International Information System for Agricultural Science and Technology) and collaborates with GFAR on the CIARD ring. He is member of the Dublin Core Advisory Board.

1 Introduction

AGROVOC is a multilingual thesaurus covering all areas of interest to the Food and Agriculture Organisation (FAO) of the UN (FAO of the UN), including agriculture, fisheries, forestry and environment. AGROVOC is now available in 19 languages, with an average of 40,000 terms in each language. AGROVOC is 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 specialised libraries, digital libraries and digital repositories to index content. It is also used as a specialised tagging resource, for the purpose of knowledge and content organisation.

First developed in the 1980's, AGROVOC has evolved over time to exploit the increasing possibilities offered by modern technologies. After its early days on paper, AGROVOC was

moved to a relational database, which represented a great improvement in terms of ease of maintenance. However, some limitations were also experienced, especially because of its distributed community of editors. Also, data was available to third parties only either by means of database dumps, or through web services. Either ways, information sharing requires a great deal of effort and control on the side of the developers and maintainers of the applications. The technologies developed within the Semantic Web approach, including the Linked Data publication style, have offered the possibility of overcoming the limitations related to the maintenance and exploitation of AGROVOC.

FAO moved to Linked Data (expressed in SKOS, Simple Knowledge Organization System), due to the advantages inherent in using a widely implemented and standard model that is both human and machine-readable. In particular, its

advantages for librarians promise to be of great value, as once thesauri are linked, the resources they index are linked as well. Also, linked data publishing offers the advantage of a single point of access using standard query languages such as SPARQL that are already widely deployed in computing applications.

This paper aims at giving a precise account of the entire process of AGROVOC maintenance and publication as Linked Data, discussing the issues we found, the lessons we learned during our work and the results we achieved. We think our work is of general interest because many thesaurus managers are embracing internet-related technologies and our work may serve as a use case to the community. We present in a single picture the current product, its past development and its social and historical context of use. As for any foundational information resource used and maintained by a geographically distributed community and exploited over the years by hundreds of different applications, innovation is not only a matter of technical research and development; it also requires careful attention to service continuity and data evolution. Therefore this paper also describes the salient aspects of publishing AGROVOC as linked data side by side with previous AGROVOC versions expressed in relational models and consumed by legacy software applications.

The rest of this paper is organised as follows. Section 2 describes the evolution of the AGROVOC formal model following the advent of the Semantic Web and the current modelling with SKOS-XL. Section 3 presents VocBench, the editing and workflow management tool for AGROVOC. Section 4 discusses the AGROVOC content maintenance, also in relation to the development of VocBench and the evolution of AGROVOC formal model. Section 5 describes the format conversion of AGROVOC into the current RDF/SKOS-XL adopted format. Section 6 presents the process followed for the generation of links between AGROVOC and ten resources relevant to the content of AGROVOC: vocabularies, thesauri and the like. Section 7 summarises and discusses the entire data flow of AGROVOC, from data maintenance to the publication as Linked Data. Section 8 describes the technical implementation of the Linked Data version of AGROVOC. Finally, in Section 9 we draw some conclusions and hint at future work.

2 Evolution of the AGROVOC model

The first attempt to bring AGROVOC to the Semantic Web dates back to 2004 (Soergel et al., 2004) and was based on the Web Ontology Language (OWL).¹ OWL was chosen because it was the best available option to move from local relational databases to the web, while allowing for a rich domain specification. As a requirement, it was assumed to stay within the OWL DL species, to benefit from the reasoning and inference capabilities of the SHOIN family (Baader et al., 2010) of description logics.

However, since thesauri are primarily terminological resources, they actually embody no notion of individual objects as opposed to classes of objects, where this distinction is actually at the basis of ontological content organisation. Also, thesaurus' contents tends to grow over time, according to different perspectives about the domain and the relations

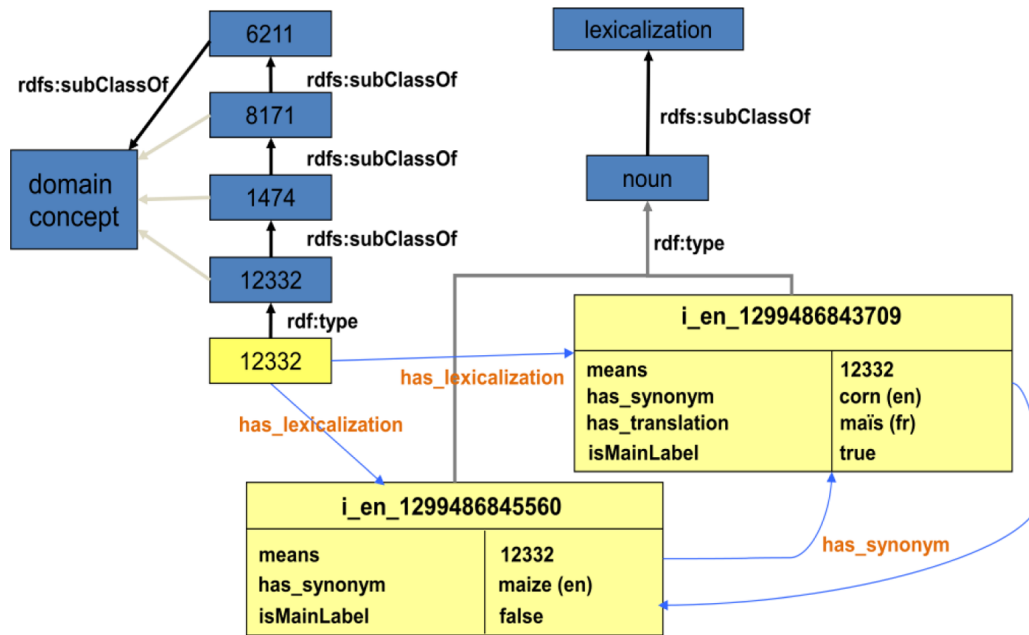
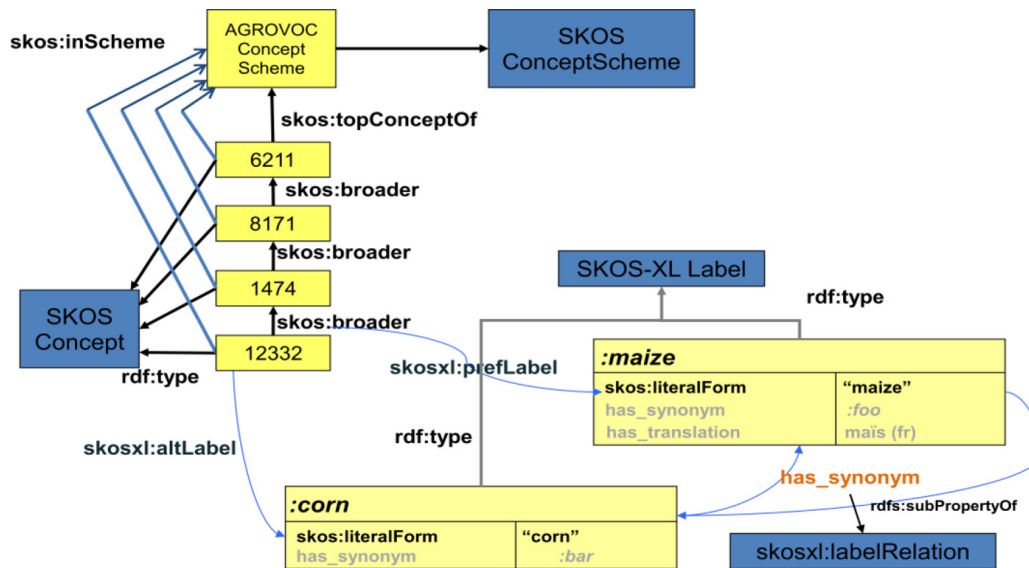
'broader than'/'narrower than' (BT/NT from now on) may be used in a variety of different situations. In fact, a thesaurus is primarily a terminological resource, hardly compatible with the rigid commitment to a logical environment required by OWL. Therefore, ad-hoc solutions had to be found in order to force the thesaurus content into the OWL metamodel. First, the requirement to stay within OWL DL (which does not allow predication over classes²), was resolved by representing domain concepts through two ontology resources: a class, organised in a hierarchy of properties `rdf:subclass` and an associated singleton instance, filled with property values.³ Second, AGROVOC is a highly multilingual resource (it is available in some 20 languages), this fact imposes requirements (Caracciolo and Sini, 2007), that are appropriately supported by OWL, i.e., by the RDF⁴ property `rdf:label` (Jupp et al., 2008). In order to be able to conveniently express terms in all the languages available in AGROVOC, yet another ad-hoc solution had to be found. A notion of lexicalisation was introduced, which forced each concept to be explicitly linked to its name, or label.

The consequences of the adopted modelling style were that the original AGROVOC hierarchy of terms was visually lost to editors and the modelling power of OWL was not exploited (because of the double hierarchy of classes and associated instances). Figure 1 shows a fragment of AGROVOC, where one concept, identified by '12332' is shown together with its associated instance and two of its names in English (maize, corn). In short, OWL was too strict to render a thesaurus resource, but at the same time it was too simplistic to model multilingual resources (for an extended discussion on this matter, also in the context of information management systems in FAO, see (Baker and Keizer, 2010).

In 2009, the W3C recommended the SKOS (W3Ca, 2009) for the rendering of resources such as thesauri over the web. As SKOS is a vocabulary for RDF specifically tailored to express thesauri, a looser semantics than that embodied by OWL is imposed on the resource. SKOS is the right choice when there is no need for formal semantics and reasoning (in particular, for classification of instances, possible in OWL thanks to the notion of object and class). Moreover, SKOS includes two properties (`skos:broader`, `skos:narrower`) to express the general thesauri relations BT/NT. In this way it is possible to directly ground relationships over concepts, whereas OWL imposes that instances must be described through properties (a constraint of the OWL DL species), while being classified through classes.

In 2009, W3C also recommended a SKOS extension for managing labels, called SKOS-XL (W3Cb, 2009). SKOS-XL offers a mechanism for treating labels (i.e., thesaurus terms) as first class objects. Labels are reified and given URIs (as opposed to being simple literals in RDF). The consequence of this approach is that with SKOS-XL, it is possible to keep track of various pieces of information about labels (e.g., date of creation and modification, editorial notes, etc.) that could not be expressed in SKOS.

In short, SKOS offers a standard vocabulary to express thesauri within RDF. With the SKOS-XL extension an appropriate linguistic characterisation of thesaurus terms can also be provided. Figure 2 shows the fragment of AGROVOC shown in Figure 1, but represented using RDF/SKOS. One can notice the use of SKOS-XL labels.

Figure 1 A fragment of AGROVOC expressed in OWL (2004) (see online version for colours)**Figure 2** A fragment of AGROVOC expressed in SKOS-XL (see online version for colours)

3 AGROVOC maintenance tool: VocBench

The storage of AGROVOC in a relational database (which is now considered a legacy system, but still in use) implied the existence of a ‘master’ copy of the database. Data would be then shipped (usually by means of SQL dump) to the editors, who would perform their editing process by means through a PHP application. Such a maintenance system was designed for use by one user at a time and did not embody any notion of editorial workflow (including change validation), which was managed informally outside the tool. The result was that no parallel work on AGROVOC was possible and the entire editorial process and above all the central control and alignment of version was error prone and time-consuming.

The great improvement of web based technologies in the years 2000 were then seen as an occasion to overcome these limitations. Moreover, given the multilingual and therefore intrinsically collaborative nature of AGROVOC, there was a need for more sophisticated functionalities than those supported by the PHP application. In particular, support was required for distributed and collaborative editing as well as for change validation within a formalised editorial workflow. Special attention to user roles and edit rights on languages was also required. A few ontology editors were already available but usually they did not support collaborative work, formalised workflow with user roles and editing rights also by languages, or UTF-8 – and none all these features together. Moreover, due to the specificity of the OWL modelling adopted at the

time, off-the-shelf ontology editing tools such as Protege (Gennari, et al., 2003; Knublauch et al., 2004) would not allow editors to graphically see the hierarchy of concepts – because the hierarchy was flattened by the use of ad-hoc properties as discussed in Section 2.

These reasons led to the development of the AGROVOC Concept Server Workbench, usually shortened into WorkBench, a web-based, fully multilingual vocabulary editor supporting distributed collaboration structured into a formalised workflow. The successor of that tool is now called VocBench.⁵ VocBench improves on its predecessor in that it fully supports a formalised workflow, by user role and by language. It supports a very fine grained mechanism of track change, in order to allow individuals and organisations to contribute to AGROVOC while maintaining the information about the provenance of their authorship. Moreover, support to multilinguality in search, visualisation and editing is fundamental to VocBench. Currently in version 1.3, VocBench – which still internally relies on the customised OWL model discussed in the previous section – is able to export data into SKOS/SKOS-XL and it will soon support these standards natively. Figure 3 presents a screenshot of the VocBench user interface showing a fragment of AGROVOC.

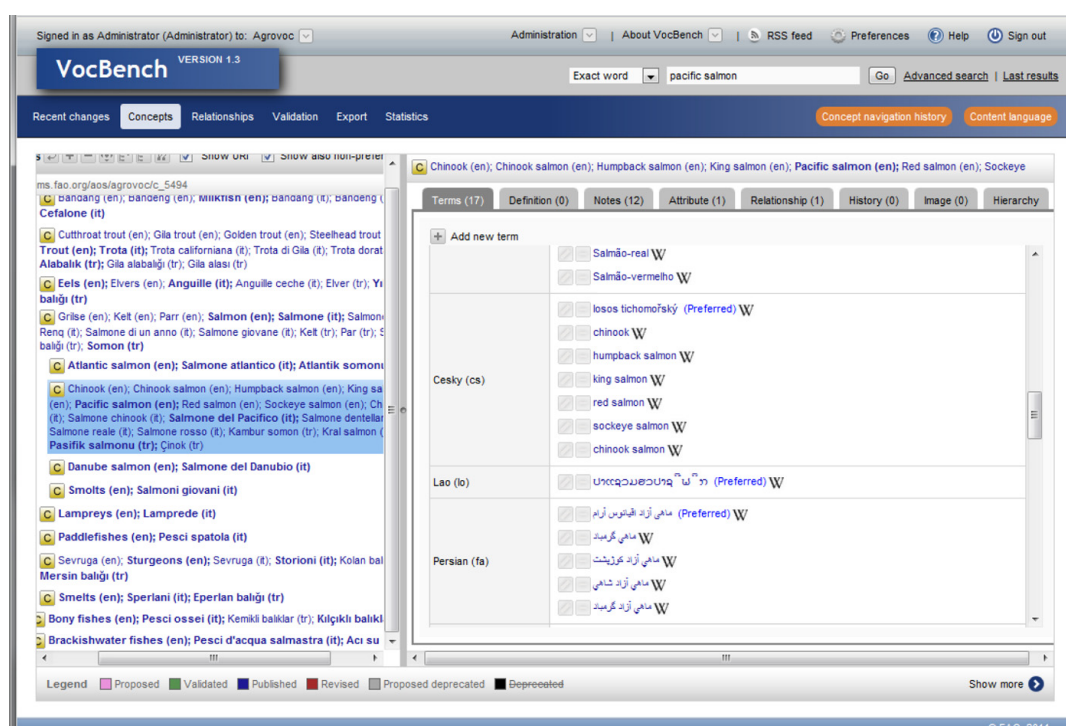
These features have made the interest around VocBench grow, which has in turn contributed to the refinement of VocBench requirements. Now VocBench is no longer an AGROVOC-only editor and its community of users has grown beyond the one originally envisaged. Currently, VocBench is used to maintain the FAO Biotechnology Glossary⁶ and much of the bibliographic metadata used by FAO.

3.1 VocBench architecture

VocBench is based on a classical three-tier architecture: presentation layer and a service layer are implemented through the Google Web Toolkit⁷ platform, while the data layer accesses an RDF triple store. In particular, the RDF access services are supported by the Protege Knowledge Management API and the data storage uses the Protege DB backend, which implements the storage of large RDF data over classical relational databases.

Currently, the data layer is being improved to better mark the separation between the abstraction layer on top of the RDF management and basic triple storage and retrieval. Such an improvement is achieved by introducing the OWL ART API.⁸ The OWL ART library provides a middle layer over different triple store technologies, so that applications exploiting its API may rely on a homogeneous and stable bus in which different, scenario-dependent technological choices can be taken. Part of the VocBench data management code has already been switched to the OWLART API through their Protege implementation, e.g., an implementation of these API appropriately translating requests in the form accepted by the Protege API. This way, all the code recently introduced and based on the OWL ART will remain stable and thus seamlessly ported to the 2.0 version, while the adopted triple store technology will probably change. The advantage of this strategy can already be seen from the tests we conduct. On the one hand, testing of VocBench on smaller portions of AGROVOC is conducted with in-memory models provided by Sesame. On the other hand, performance and scalability tests are conducted on high performance triple stores. At the same time, transaction-based triple stores with

Figure 3 VocBench v1.3. User interface showing a fragment of AGROVOC (see online version for colours)



an optimum tradeoff between efficiency and scalability, will act as backend for future versions of VocBench. Similarly to the Jena API (McBride, 2001) and to the Manchester OWL API (Bechhofer et al., 2003), OWLART also features high level access methods specifically tailored for the various vocabularies of the RDF family. Currently supported vocabularies are RDF, RDFS, OWL (1st version), SKOS and SKOS-XL. These vocabulary APIs hide most of the triple management and provide abstract methods tightly connected with the specific RDF interpretation: for instance, in SKOS, the related vocabulary API manage much of the work that is necessary in order to avoid breaking the modelling constraints expressed in the SKOS specifications (but which are not formally expressed in the SKOS model itself, thus requiring external support by dedicated machinery).

Support for OWL ontologies is also on the roadmap for future VocBench releases. As noted earlier, OWL is useful when a clear distinction between individual and classes is needed. For example, this is the case of authority files for journals and other bibliographic data. This is in fact the next type of resource in line to be maintained through VocBench.

Given that VocBench still internally relies on the legacy OWL model for AGROVOC (see Section 2), its native format is not suitable for linked data publication as-is. Periodical conversions are thus made towards the SKOS-XL format for LOD publication (see Section 7 for a detailed description of the AGROVOC maintenance lifecycle).

4 AGROVOC content and maintenance

AGROVOC grew over the years, both in terms of its content and in terms of the languages in which its contents is available. Originally created in 3 languages, English, French, Spanish, it is now published in 19 languages and 6 more are under development. For many years it was stored in a relational database, with the implications in editorial maintenance that we described in the previous section. Also, that implied a quite rigid, centralised and e-mail based communication between FAO and the groups editing the various languages of AGROVOC. With the move to modern, web-based technologies (see Sections 2 and 4), it was expected that the management of AGROVOC would be streamlined and so the communication with/between AGROVOC editors. VocBench is now in use and we can see its effect on the management of AGROVOC content. First, it gave new impulse to the translation of AGROVOC that were in progress. Second, new translations are now in progress using VocBench. The effects also outcome our expectation, as it was not planned to have a general editor for other resources.

The revision of the AGROVOC model was also taken as an occasion to revise its content from a structural point of view. The AGROVOC structure was reorganised so as to reduce the number of top concept from some hundreds to 25. Also, a number of non-hierarchical relations were introduced, that are now under revision in order to harmonise them to AGROVOC current formal modelling. In summary, the adoption of SKOS-XL turned out as an occasion for enrichment of AGROVOC terminological content and

rationalisation of its NT/BT structure. Also, the requirement analysis for the development of VocBench contributed to the refinement of what information should be attached to AGROVOC concepts. For example, detailed information about authorship and history of change is needed within the workflow and it is expected to have positive impact on the collaboration among AGROVOC editors.

5 From VocBench internal model to SKOS-XL

As previously explained in Section 2, SKOS-XL is used for publishing AGROVOC as linked data, while VocBench still internally relies on the (customised) OWL-based model for AGROVOC that we discussed (see Section 2). Given that this internal data model will be in use until a fully SKOS-compliant release of VocBench is developed, a conversion process is needed in order to make AGROVOC easily available as linked data.

The conversion is performed by exploring AGROVOC concept by concept (by navigating the tree of domain concepts) and then properly converting all associated elements (the class realising the concept in the tree, the associated singleton instance realising the concept as an editable object and its relationships). An alternative approach would have been to perform a triple-by-triple based conversion, which was avoided for two main reasons:

- 1 Conversion is not based on a 1-to-1 translation of predicates: the port to the SKOS-XL model described in section 2 implies that any given source predicate may not always be translated to the same predicate from the target vocabulary and this translation depends instead on the context of the application of the predicate (thus, subject and object of the triple featuring that predicate, where the nature of these subject and object is explicated in other triples). Complex transformations involving patterns of several triples have thus been made necessary in some cases, where the misalignment between the two models goes far beyond terminological issues
- 2 VocBench internally uses the Protégé OWL API (Knublauch et al., 2004) backed by the Protégé DB, which does not allow for easy processing of triples. The Protege DB (which allows for storage of Protégé resources over a relational database) uses an extension of the old Protégé Frame model as an inner model, which is based on a purely object-oriented paradigm and is not based on triples. The difficulty in a triple-by-triple conversion lies in this model, which uses different ‘bags’ for classes, instances and properties. Their role is not inferred by their position in RDF triples, but by their explicit membership to one of these bags. For this reason, Protégé does not allow for full support for querying triples and mostly relies on a live-export of the model as a Jena read-only triple store. This export mechanism is not reliable when used with the Protégé DB backend and is very slow for very large repositories (as the Jena model is recreated in-memory), so the conversion process natively uses Protégé’s API to access AGROVOC resources.

To summarise the process, the Protege API (with DB backend) are used to read the legacy OWL version of the data and the OWLART API (by adopting the SKOSXL Model interface of OWLART and the Sesame2 (Broekstra et al., 2002) implementation for the API) is used to convert the data versus the target SKOS-XL model and exported in NTRIPLES and RDFXML files, which are then used for linked data publication.

6 Linking AGROVOC to other resources

We started the enterprise of linking AGROVOC to other resources with the expectation that by linking thesauri and vocabularies also the information (e.g., data repositories) attached to them will be available.⁹ Figure 4 depicts an intuitive view of the benefit of linking AGROVOC to other resources.

AGROVOC entered the linked data cloud with links to ten resources, vocabularies, thesauri and ontologies in areas related to domain covered by AGROVOC. Six of the linked resources are of general coverage: the Library of Congress Subject Headings (LCSH)¹⁰, the NAL Thesaurus¹¹, RAMEAU Repertoire d'autorite-matiere encyclopedique et alphabetique unifie¹², Eurovoc¹³, DBpedia¹⁴ and an experimental Linked Data version of the Dewey Decimal Classification.¹⁵ The remaining four resources are specific to various areas of interest: GEMET¹⁶ is specialised on environment, the STW Thesaurus for Economics¹⁷ covers the domain of economy, the SOZ Thesaurus¹⁸ is about social science and the FAO Geopolitical Ontology¹⁹ is an ontology about countries and political regions.²⁰ The linked resources are mostly thesauri,

already available as RDF/SKOS resources, relevant to the domains covered by agriculture, forestry, fisheries, food and geographic. The RDF/SKOS version of other resources, such as ASFA, Biotech Glossary (FAO), etc., are in progress.

The criteria for selecting those resources were:

- their coverage
- the type of informative resources associated to them
- their availability as Linked Data, or RDF/SKOS format.

All data repositories considered for alignment with AGROVOC are available as SKOS or RDF and we were able to load them on a local triple-store (in this case, Sesame²¹). Thesauri were considered in their entirety except RAMEAU, for which only agriculture related concepts were considered (amounting to some 10% of its 150 thousand concepts). Candidate mappings were found by applying string similarity matching algorithms to pairs of preferred labels (Cohen et al., 2003)²² and by exploiting Ontology Alignment API (Euzenat, 2004) for managing the produced matchings. During the process only one common language of the two resources was considered as the matching methods used did not support more than one language label at a time. The single language in common was English in all cases except one, as AGROVOC and RAMEAU only have French in common.

Table 1 shows, for each resource linked to AGROVOC, its area of coverage (second column), the language considered for mapping with AGROVOC (third column) and the number of exact matches resulting from the evaluation (fourth columns).

Figure 4 An intuitive view of the benefit of linking various thesauri together (see online version for colours)

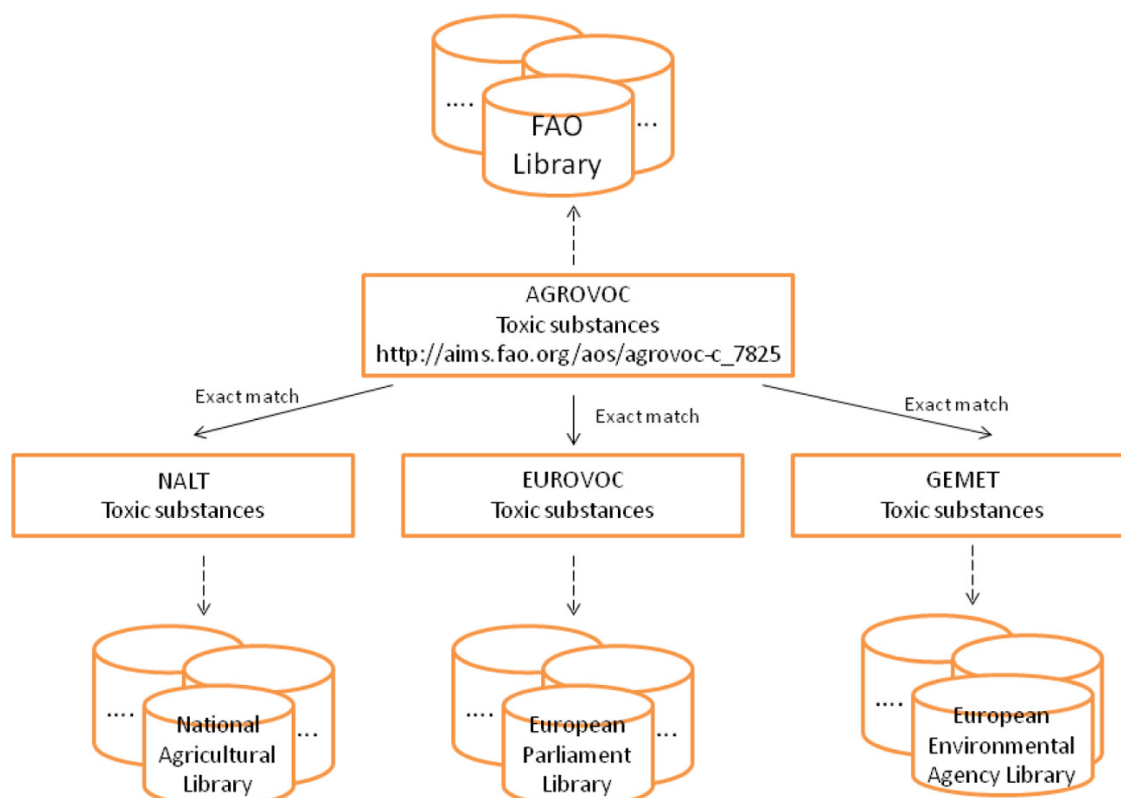


Table 1 Resources linked to AGROVOC

<i>Vocabulary</i>	<i>Coverage</i>	<i>Languages considered</i>	<i># skos:exact Match</i>
EUROVOC	General	EN	1297
DDC	General	EN	409
LCSH	General	EN	1093
NALT	Agriculture	EN	13,390
RAMEAU	General (cut on Agri.)	FR	686
DBpedia	General	EN	1099
TheSoz	Social science	EN	846
STW	Economy	EN	1136
FAO	Geopolitical	EN	253
Geopolitical Ontology	information		
GEMET	Environment	EN	1,191

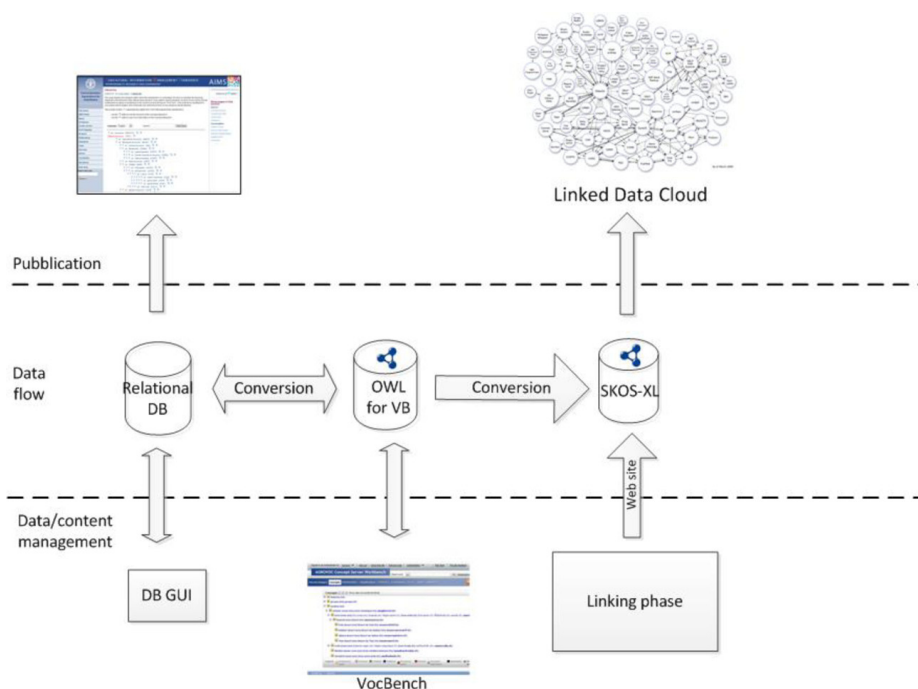
Candidate links were presented to a domain expert for evaluation in the form of a spreadsheet. Once validated the mappings were loaded in the same triple store where the linked data version of AGROVOC is stored. All resulting validated candidate matching were considered of the type `skos:exactMatch`.

Our objective when linking AGROVOC to other resources was to provide only main anchors and privilege accuracy over recall of potential links. This is the reason 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 (Gale et al., 1992) supports our claim that case, similar strings correspond to equivalent meanings. The use of more sophisticated context-based approaches might have contributed to filtering out the potential results, more than widening their number (thus incrementing precision over recall), however this potential loss of precision was much compensated by the manual validation of candidate links done by a domain expert.

7 AGROVOC LOD data flow

Figure 5 provides a high-level view of the entire AGROVOC maintenance process and of 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). One may notice that two maintenance tools are shown in the bottom layer and three data formats/repositories are shown in the middle layer.

The relational database is still in use as it serves as a master repository of AGROVOC for many existing applications. Also, it is kept because some editors are still not able to adopt VocBench. In some cases, this is due to scarce or unreliable bandwidth, in other cases this is due to the preference of editors well acquainted with old tool. To address the former case, we are considering making available local installations of VocBench with batch inclusion in the master copy, while to address the latter case we are designing training resources and guidelines specifically for editors. One may notice the cycle of conversion between the relational format and the OWL format. Such a conversion is needed to synchronise the data accessed by editors using the two data maintenance tools currently in use. This duplication of data repository and consequent data conversions is obviously not ideal and in principle it should be limited as much as possible. Since its first appearance in 1980s, AGROVOC has supported a worldwide community of users (people and institutions), who have developed a number of applications relying on the legacy relational model. These applications require support and so some of these conversion steps are unavoidable. This setting gives an idea about the complexity of a scenario where a notable resource which has been made publicly available for years (both online and, indirectly, through the many applications which access it) is migrated to a new standard. Elaborated procedures are made necessary and the conversion effort, modelling issues and

Figure 5 Overview of the process for publishing AGROVOC as linked data (see online version for colours)

information are just the tip of the iceberg composed of the real effort spent in maintaining it and all the services targeted towards its legacy format(s).

One may also notice that the link generation phase is currently external to the traditional AGROVOC editorial cycle and added to the converted SKOS-XL repository. This is the reason why Figure 5 does not show any arrow connecting the web site publication and the linked data publication.

Several conversion steps are then present in the AGROVOC lifecycle. Two formats are used for maintenance and a third format (more in line with current semantic standards for representing thesauri in the Semantic Web) is used for publishing AGROVOC contents as Linked Open Data. Contributions to AGROVOC may come from tools following any of the available formats and not only by the main authoring tool VocBench – see the bidirectional arrow of conversion shown in Figure 5. Any single version of AGROVOC, say version ‘i’, is edited both through the relational database and VocBench. A new version then, will contain the merging of all changes coming from the two editing lines. In particular, the users of the relational database (or of local copies of VocBench) send their contents to FAO. Content is then merged (i.e., merging of data coming from web-based VocBench and relational DB copies. At some point, also data coming from local copies of VocBench will have to be used) following different strategies. The result is used to produce version ‘i + 1’.

The merging strategies may vary in the applied methodology and related complexity, depending on the kind of contribution which need to be included. Usually, addition of labels for a new language is a pretty straightforward procedure: whether it has been applied to the relational database or to the VocBench version, a simple addition of all produced data is necessary, possibly mediated by conversion from one format to the other. Even addition of new mapping relationships between



AGROVOC and other LOD thesauri and resources in general is not demanding upon the mere task of bringing them in the main development trunk (e.g., without considering the effort to produce and validate them). Conversely, changing even few URI names or, even worse, applying even slight modifications to any naming policy for an entity type (e.g., URI for the labels, which are reified inside the thesaurus) is a critical operation which needs to be operated carefully, to avoid misalignments between the various parallel realisations coexisting in a same version of AGROVOC.

So far, when a VocBench version is finalised with contributions coming from different sources and according to different formats, it is then converted back to relational DB and used for applications based on it as well as for users editing AGROVOC through legacy applications. At the same time, a SKOS-XL version is produced from the VocBench one and enriched with information, such as metadata descriptors from the void vocabulary²³ to feed the LOD endpoint with updated data.

8 AGROVOC as linked data

The linked data version of AGROVOC is now available online thank to a collaboration between FAO and MIMOS Berhad.²⁴ Data is stored in an RDF triple store (Allegrograph²⁵) 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 version of Pubby²⁶ hosted on FAO servers with customised Velocity Templates²⁷ (to provide more readable labels for properties in some cases, hide redundant data, etc.). Figure 6 shows an AGROVOC concept in the HTML visualised we provide. The HTML representation of

Figure 6 An HTML representation of an AGROVOC concept as linked data (see online version for colours)

 maize at AGROVOC Thesaurus http://aims.fao.org/aos/agrovoc/c_12332 	
Property	Value
altLabel	<ul style="list-style-type: none"> Granoturco (it) Ziarno kukurydzy (pl) corn (maize) (en) kukurica (hu) zmo (kukurica) (sk) zmo kukurice (cs) зерно кукурузы (ru) अनाज (मक्का) (hi) コーン (ja) 苞谷 (zh)
broader	<ul style="list-style-type: none"> <http://aims.fao.org/aos/agrovoc/c_1474>
closeMatch	<ul style="list-style-type: none"> <http://dbpedia.org/resource/Maize>
exactMatch	<ul style="list-style-type: none"> <http://d-nb.info/gnd/4037135-9> <http://eurovoc.europa.eu/1744> <http://zbw.eu/stw/descriptor/14093-4>
date of creation	1981-01-26
date of last update	2005-08-09
isProducedBy	<ul style="list-style-type: none"> <http://aims.fao.org/aos/agrovoc/c_8504>
isUsedToMake	<ul style="list-style-type: none"> <http://aims.fao.org/aos/agrovoc/c_15500> <http://aims.fao.org/aos/agrovoc/c_25472>
narrower	<ul style="list-style-type: none"> <http://aims.fao.org/aos/agrovoc/c_14385> <http://aims.fao.org/aos/agrovoc/c_2187> <http://aims.fao.org/aos/agrovoc/c_2974> <http://aims.fao.org/aos/agrovoc/c_6108> <http://aims.fao.org/aos/agrovoc/c_7152> <http://aims.fao.org/aos/agrovoc/c_7552> <http://aims.fao.org/aos/agrovoc/c_8337>
prefLabel	<ul style="list-style-type: none"> Kukurydza (ziarno) (pl) MAIS (de) Mais (it) Maiz (es) Mais (fr) Milho (pt)

AGROVOC as linked data is accessible from: <http://aims.fao.org/standards/agrovoc/linked-open-data>.

Both RDF and HTML accesses are resolved through content negotiation (and redirection to MIMOS where appropriate) on the FAO servers, to expose the FAO domain on the LOD cloud.

9 Conclusions

AGROVOC's maintenance, alignment with other thesauri and publication as linked data is supported by an entire publishing chain, consisting of users engaged in a workflow supported by specialised tools. In particular, the remodelling of AGROVOC using OWL and SKOS and its publication as linked data implies 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 visualisers such as Pubby and 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 maintenance process, both historical information management systems and new semantically-aware systems play a role. A sequence of conversion steps, some of which could in principle be streamlined, is not ideal. But support for previous versions and their user base is a business process requirement that cannot be ignored. Work is ongoing to provide training to AGROVOC editors, organise workshops for data managers and improve VocBench functionalities so that it can be used by all. Also, the quality control of AGROVOC content (for both its terminological and structural aspects) is a continuous activity.

In this light, the immediate issues to address include the creation of an easy to use and install version of VocBench that can be used off-line and synchronized with the central repository (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 standardisation 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 organising their data in SKOS. Another planned development is the integration within VocBench of the alignment functionalities that are currently hosted in Eclipse and used to extract and validate links to other resources. This will integrate the alignment workflow with the overall AGROVOC editing workflow. From the content point of view, we plan to continue linking AGROVOC to other resources and to start using `skos:closeMatch` in addition to `skos:exactMatch`.

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.

Acknowledgements

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, Jerome Euzenat, Stefan Jensen, Antoine Isaac, Soren Roug, Thomas Baker and Mary Redahan.

References

- Baader, F., Calvanese, D., McGuinness, D.L., Nardi, D. and Patel-Schneider, P.F. (Eds.) (2010) *The Description Logic Handbook: Theory, Implementation and Applications*, 2nd ed., Cambridge University Press, Cambridge.
- Baker, T. and Keizer, J. (2010) 'Linked data for fighting global hunger: experiences in setting standards for agricultural information management', in Wood, D. and Wood, D. (Ed.): *Linking Enterprise Data*, Springer-Verlag New York Inc., Washington, DC USA, Vol. 4, pp.177–201.
- Bechhofer, S., Lord, P. and Volz, R. (2003) 'Cooking the semantic web with the OWL API', *2nd International Semantic Web Conference, ISWC*, Sanibel Island, Florida.
- Broekstra, J., Kampman, A. and van Harmelen, F. (2002) 'Sesame: a generic architecture for storing and querying RDF and RDF schema', *The Semantic Web – ISWC 2002: First International Semantic Web Conference*, pp.54–68, Springer Berlin/Heidelberg, Sardinia, Italy.
- Caracciolo, C. and Sini, M. (2007) 'Requirements for the treatment of multilinguality in ontologies within FAO', *Proceedings of OWLED 2007*, CEUR workshop proceedings. <http://ceur-ws.org/Vol-258/>
- Cohen, W.W., Ravikumar, P. and Fienberg, S.E. (2003) 'A comparison of string distance metrics for name-matching tasks', *IJCAI-2003*, pp.73–78.
- Euzenat, J. (2004) 'An API for ontology alignment', in McIlraith, S.A., Plexousakis, D. and van Harmelen, F. (Ed.): *The Semantic Web – ISWC 2004: Third International Semantic Web Conference*. 3298, Springer, Hiroshima, Japan, pp.698–712.
- Gale, W., Church, K. and Yarowsky, D. (1992) 'A method for disambiguating word senses in a large corpus', *Computers and the Humanities*, Vol. 26, pp.415–439.
- Gennari, J., Musen, M., Fergerson, R., Grosso, W., Crubezy, M., Eriksson, H., et al. (2003) 'The evolution of Protege-2000: an environment for knowledge-based systems development', *International Journal of Human-Computer Studies*, Vol. 58, No. 1, pp.89–123.
- Jupp, S., Bechhofer, S. and Stevens, R. (2008) SKOS with OWL: don't be full-ish!', in Dolbear, C., Ruttenberg, A. and Sattler, U. (A cura di): *OWLED*, p.432, CEUR-WS.org
- Knublauch, H., Fergerson, R.W., Friedman Noy, N. and Musen, M.A. (2004) 'The protege OWL plugin: an open development environment for semantic web applications', *Third International Semantic Web Conference – ISWC 2004*, Hiroshima, Japan.
- McBride, B. (2001) *Jena: Implementing the RDF Model and Syntax Specification*, Semantic Web Workshop, WWW2001.

Morshed, A., Caracciolo, C., Gudrun, J. and Keizer, J. (2011) 'Thesaurus alignment for linked data publishing', *Proc. of Dublin Core 2011 (forthcoming)*, <http://dcpapers.dublincore.org/index.php/pubs/issue/view/163>

Soergel, D., Lauser, B., Liang, A., Fisseha, F., Keizer, J. and Katz, S. (2004) 'Reengineering thesauri for new applications: the AGROVOC example', *Journal of Digital Information – JODI*, Vol. 4, <http://eprints.rclis.org/handle/10760/15694?mode=full>

W3Ca (2009 18-August) *SKOS Simple Knowledge Organization System eXtension for Labels (SKOS-XL)*, Miles, A. and Bechhofer, S. (Eds.): Retrieved 2011 22-March from World Wide Web Consortium (W3C), <http://www.w3.org/TR/skos-reference/skos-xl.html>

W3Cb (2009 18-August) *SKOS Simple Knowledge Organization System Reference*, Miles, A. and Bechhofer, S. (Eds.): Retrieved 2011 22-March from World Wide Web Consortium (W3C), <http://www.w3.org/TR/skos-reference/>

Notes

¹<http://www.w3.org/TR/owl-ref/>

²Ontology classes are first order objects (i.e., *predicates*). The assignment of values to a property of a class is equivalent to predicate over a predicate, that is, to move to an higher order logic. The result is not OWL DL.

³Interestingly, the concept/object dualism was resolved in second version of the OWL language in a similar way, with a reasoning expedient called *punning*. At inference level, punning splits a unique URI reference into two objects implementing its dual nature of individual and class. See http://www.w3.org/TR/owl2-new-features/#F12:_Punning.

⁴<http://www.w3.org/TR/rdf-schema/>

⁵<http://aims.fao.org/tools/vocbench-2>

⁶<http://www.fao.org/biotech/biotech-glossary/en/>

⁷<http://code.google.com/webtoolkit/>

⁸<http://art.uniroma2.it/owlart/>

⁹The W3C maintains a page collecting information about web applications built on top of linked data: <http://www.w3.org/wiki/SweoIG/TaskForces/CommunityProjects/LinkingOpenData/Applications>

¹⁰<http://id.loc.gov/authorities/subjects.html>

¹¹<http://agclass.nal.usda.gov/>

¹²<http://rameau.bnf.fr/informations/rameauenbref.htm>

¹³<http://eurovoc.europa.eu/>

¹⁴<http://dbpedia.org/About>

¹⁵<http://dewey.info/>

¹⁶<http://www.eionet.europa.eu/gemet>

¹⁷<http://zbw.eu/stw/versions/latest/about>

¹⁸<http://www.gesis.org/en/services/tools-standards/social-science-thesaurus/>

¹⁹<http://www.fao.org/countryprofiles/geoinfo.asp>

²⁰For an updated list of resources linked to AGROVOC, see <http://aims.fao.org/standards/agrovoc/linked-open-data>

²¹<http://www.openrdf.org/>

²²<http://alignapi.gforge.inria.fr/>

²³<http://www.w3.org/TR/void/>

²⁴<http://www.mimos.my/>

²⁵<http://www.franz.com/agraph/allegrograph/>

²⁶<http://www4.wiwiwiss.fu-berlin.de/pubby/>

²⁷<http://velocity.apache.org/>