Thesaurus Maintenance, Alignment and Publication as Linked Data

The AGROVOC Use Case

Caterina Caracciolo[*], Ahsan Morshed[*], Armando Stellato[+], 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 Roma, Italy
{caterina.caracciolo,ahsan.morshed,
gudrun.johannsen,yves.jaques,johannes.keizer}@fao.org

[+]ART Group, Dept. of Computer Science, Systems and Production
University of Rome, Tor Vergata
Via del Politecnico 1, 00133 Rome, Italy
stellato@info.uniroma2.it

Abstract. The AGROVOC multilingual thesaurus maintained by the Food and Agriculture Organization of the United Nations (FAO) is now published as linked data. In order to reach this goal AGROVOC was expressed in Simple Knowledge Organization System (SKOS), and its concepts provided with dereferenceable URIs. AGROVOC is now aligned with ten other multilingual knowledge organization systems 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.

Keywords. AGROVOC; Mapping; Agriculture; linked data

1 Introduction

AGROVOC is a multilingual thesaurus covering all areas of interest to the Food and Agriculture Organization of the UN (FAO of the UN), including agriculture, fisheries, forestry, environment, etc. First developed in the 1980's, 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. It is

used worldwide by researchers, librarians, and information managers for indexing, retrieving, and organizing data in agricultural information systems.

FAO moved to linked data expressed in SKOS 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 presents the result of this work and the process followed to achieve it. It presents in a single picture the current product, its past development, and its social and historical use context. 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 organized as follows. Section 2 describes the evolution of the AGROVOC model and content following the advent of the Semantic Web. Section 3 presents VocBench, the editing and workflow management tool for AGROVOC. Section 4 concerns itself with the conversion of AGROVOC into an RDF/SKOS-XL resource. Section 5 is about publishing AGROVOC as linked data, and Section 6 presents the process followed to generate candidate links from AGROVOC to other thesauri. Section 7 summarizes and discusses the entire process of generating a linked data version of AGROVOC. Section 8 concludes.

2 Evolution of the AGROVOC model and content

The first attempt to bring AGROVOC to the Semantic Web dates to 2004 [1], and was based on Ontology Web Language (OWL). OWL was chosen because it allows for rich domain specification in its distinction between objects and classes of objects. However, as thesauri do not recognize a difference between object and class some forcing was made, which in turn made it problematic to use editing tools such as Protégé [2,3]. Fig. 1 provides a sketch of how AGROVOC content was organized in an OWL model. In that model, concepts were organized in a hierarchy defined through the classical `rdfs:subClassOf`. Thesaurus relations "broader term" and "narrower term" (BT/RT) were rendered by means of ad hoc OWL object properties, and their properties were attached to singleton instances of each class, which represented the concept itself. So, actually, each concept was represented though two resources: a class, organized in the hierarchy, and its associated singleton instance, filled with property values. This choice was made to remain inside the boundaries of OWL DL. Also, labels were managed by introducing a notion of lexicalization, which forced each concept to be explicitly linked to its name, or label. The consequences of

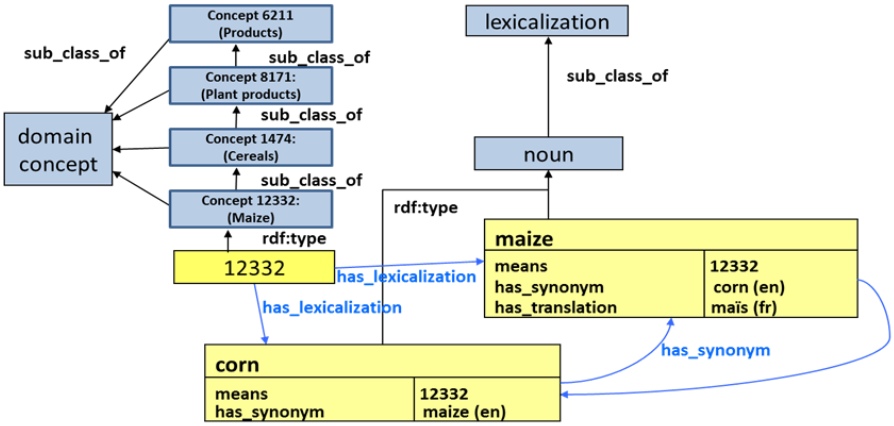

Fig. 1. AGROVOC Legacy Model based on OWL (2004)

this modeling style were that the original AGROVOC hierarchy of terms was visually lost to editors, while the modeling power of OWL was not exploited. In short, OWL was too strict to render a thesaurus resource, but at the same time it was too simplistic to model multilingual resources.

In 2009, the W3C recommended the Simple Knowledge Organization System (SKOS) [4] 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 the same year, W3C also recommended a SKOS extension for managing labels, called SKOS-XL [5]. 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 characterization of thesaurus terms can also be provided. This is the reason why the previous attempt to express AGROVOC in OWL has been superseded by a SKOS-XL modeling.

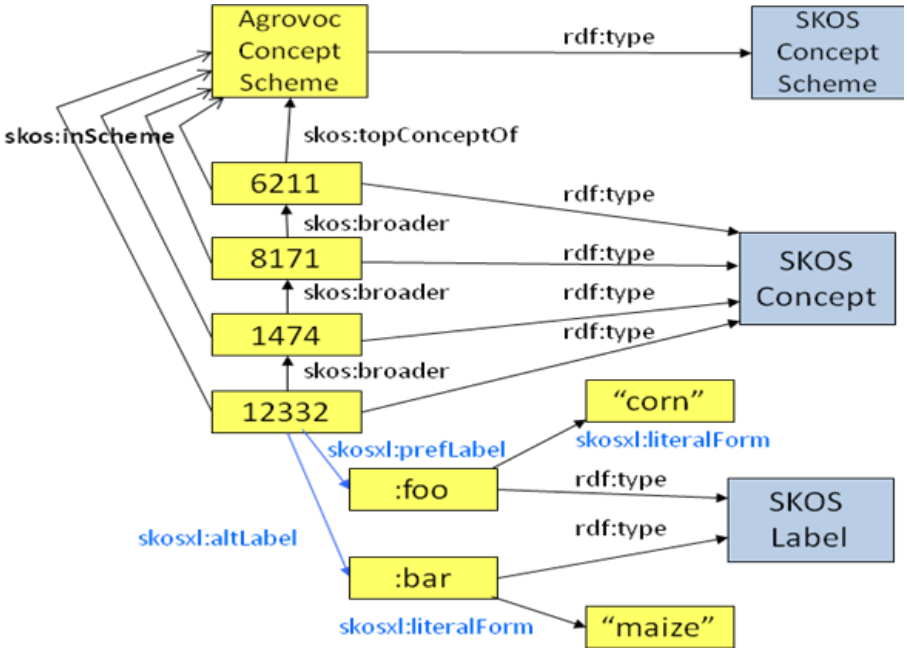

Fig. 2. AGROVOC Current Model (plain SKOS-XL)

In parallel with the definition of the most appropriate formal representation of AGROVOC for web consumption, AGROVOC also underwent a massive revision of its content. The number of top concepts was reduced to 25, and the hierarchies were reorganized accordingly. Also, a number of domain specific relations between concepts were added, defined globally for the entire AGROVOC. A future improvement in this direction is to simplify where possible, and standardize the domain specific relations introduced.

3 Support for AGROVOC editorial maintenance: VocBench[1]

From its inception, the AGROVOC thesaurus was stored in a relational format. In its relational model, AGROVOC was treated as a purely terminological resource, with no notion of concepts. Local identifiers were used to connect terms used in different languages to express the same meaning. Data maintenance was possible through a web application, developed in PHP and connected to the master database. Such a maintenance system was designed for use by one user at a time, and did not embody

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

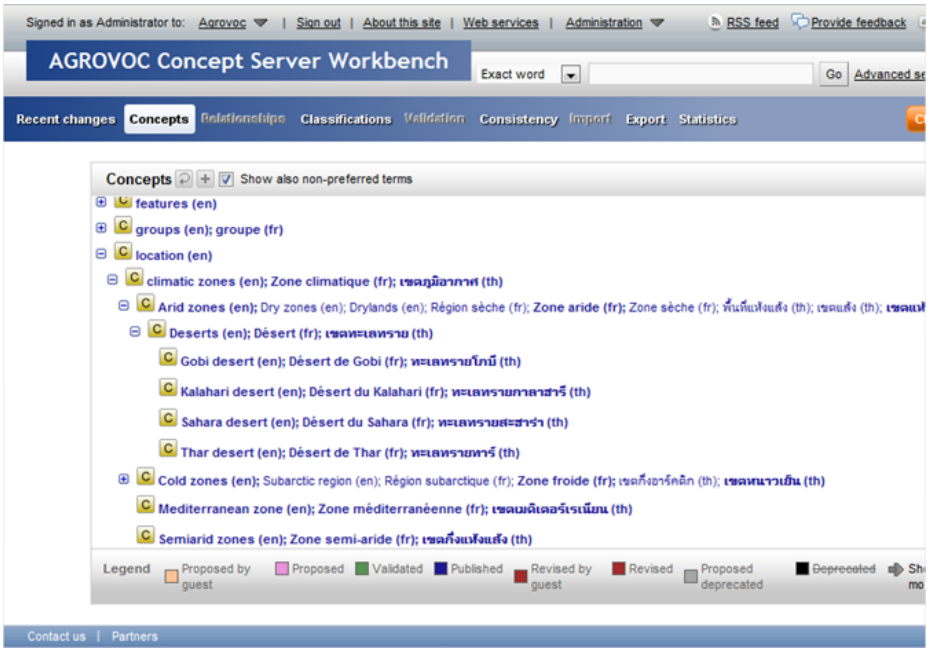

Fig. 3. User interface of VocBench v1.1, visualizing a fragment of AGROVOC

any notion of editorial workflow (including change validation), which was managed informally outside the tool.[2]

With AGROVOC's shift to the Semantic Web, the need emerged for an adequate way to manage its content. Due to the specificity of the OWL modeling adopted at the time, the use of traditional ontology editing tools (e.g., Protégé) was cumbersome. 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 now required for distributed and collaborative editing as well as change validation within a formalized editorial workflow. Special attention to user roles and edit rights on languages was also required. This led to the development of the AGROVOC Concept Server Workbench, a web application meant to serve as the web-based platform for AGROVOC maintenance.

When AGROVOC was re-modeled in SKOS, the AGROVOC Concept Server Workbench followed, evolving into a general purpose (i.e., no longer exclusively AGROVOC based), SKOS-compliant platform for collaborative knowledge management. It was thus renamed VocBench.

Fig. **3** presents a screenshot of the VocBench user interface showing a fragment of AGROVOC.[3] VocBench improves on its predecessor in that it fully supports a

[2] Note that that tool is still in use, as discussed in Sec. 8.

[3] At the time of writing VocBench is released as version 1.1, while version 1.2 is in phase of beta testing.

formalized workflow, by user role and by language. Moreover, functionalities related to change tracking, translation of element names, and search across/within languages are fundamental to VocBench. VocBench still internally relies on the customized OWL model discussed in the previous section. However, it also enables data import and export to SKOS/SKOS-XL, which makes the application usable with other data sets.[4]

In order to improve its generality, VocBench's next major release (2.0) will feature a native interface for SKOS and SKOS-XL based on the OWL ART API[5] abstraction layer and SKOS-XL interfaces. This library for RDF 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. As an example, part of the current VocBench has already been switched to the OWLART API through its Protégé wrapper, code which will remain stable and thus seamlessly ported to the 2.0 version. At the same time, testing of VocBench on smaller portions of AGROVOC is conducted with in-memory models provided by Sesame, while performance and scalability tests are conducted on high performance triple stores (which will probably back the deployed VocBench). OWLART also features – as for the Jena API [6] and the Manchester OWL API [7] used in Protégé 4 – 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 they manage much of the work which is necessary in order to avoid breaking the formal modeling constraints expressed in the specifications.

Support for generic 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, as in the case of the FAO Journal Authority Data Collection (JAD), the next in line to be maintained through VocBench.

Given that VocBench still internally relies on the customized OWL model for AGROVOC, its native format is not suitable for linked data publication as-is. Periodical conversions are made into SKOS-XL format.

4 Conversion from VocBench internal model into SKOS-XL

As previously noted, SKOS-XL is used for publishing AGROVOC as linked data, while VocBench still relies internally on the legacy customized-OWL model for AGROVOC. 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 concept tree) and then properly converting all associated elements (the

[4] Internally to FAO, VocBench is also used for the management of the Biotech Glossary. See http://www.fao.org/biotech/biotech-glossary/en/.

[5] http://art.uniroma2.it/owlart/

class realizing the concept in the tree, the associated singleton instance realizing the concept as an editable object, and its relationships). Another possible approach would be to perform a triple-by-triple based conversion, which was avoided because:

1. According to the Model translation directives: the same predicate may not always be translated the same way, but depends on its context (subject and object)
2. VocBench internally uses the Protégé API [2] backed by the Protégé DB, which does not allow for easy processing of triples. The Protégé 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. 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 role in RDF triples, but by their explicit membership to one of these bags. For this reason, Protégé does not allow an easy processing of triples, and mostly relies on a live-export of the model as a Jena read-only triple store. This export is known to be problematic so the conversion process natively uses Protégé's API to access AGROVOC resources.

To summarize the process, the Protégé API (with DB backend) are used to read the legacy OWL version of the data and the OWLART API (by adopting the `SKOSXLModel` interface and the Sesame2 [8] implementation for the API) is used to convert the data in an NTRIPLES and RDFXML file, which is then used for linked data publication.

5 Technical setup of publishing AGROVOC as linked data

The linked data version of AGROVOC is now available online owing to collaboration between FAO and MIMOS Berhad[6]. Data is stored in an RDF triple store (Allegrograph[7]) hosted on a high-performance server in Kuala Lumpur. A SPARQL endpoint, combined with an http resolution of its entities, allows for publication as linked data. The HTML representation of linked data is made available through a version of Pubby[8] with customized velocity templates, providing more readable labels for properties in some cases, hiding redundant data, etc. As an example of the human readable visualization of an AGROVOC concept in linked data, see http://aims.fao.org/aos/agrovoc/c_330892.

6 Linking AGROVOC to other Resources

AGROVOC entered the linked data cloud with links to some ten resources (mostly thesauri, already available as RDF/SKOS resources, some of them also published as

[6] http://www.mimos.my/
[7] http://www.franz.com/agraph/allegrograph/
[8] http://www4.wiwiss.fu-berlin.de/pubby/

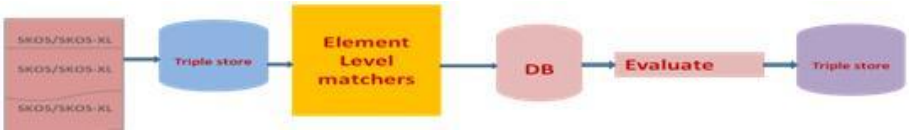

Fig. 4. Matching Process

linked data) relevant to the domains covered by AGROVOC. Others are in progress[9]. This section describes the process adopted to identify those links: see Fig. **4** for a schematic view of the process. A detailed description of the process of providing AGROVOC with links to other thesauri, in the linked data style, can be found in [9]

All data repositories considered for alignment with AGROVOC are available as SKOS-RDF, and could be loaded in a local triple-store (in this case Sesame[10]).[11] All possible pairs of concepts were considered, where the first concept in the pair comes from AGROVOC, and the second concept comes from one of the other thesauri. For each of the pairs of concepts so extracted, one preferred label per concept was selected (for the language being matched) and string similarity measures between labels was applied. Note that in this process only preferred labels in one language were 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, where French was the common language.

A selection of the most common string similarity measures was used [10], as implemented in the Alignment API[12] [11]. In order to combine these similarity values into a single number, an arithmetic average of all similarity values was computed, which seemed appropriate for a first attempt. Finally, an empirically identified threshold was applied to select candidate matches for further evaluation.

The candidate matches 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. This allows AGROVOC data to also display its outbound links in the style of linked data publishing[13].

[9] For an updated list of resources linked to AGROVOC, see http://aims.fao.org/standards/agrovoc/linked-open-data

[10] http://www.openrdf.org/

[11] The entire thesauri were considered in all cases except in the case of RAMEAU, for which agriculture related concepts were considered (amounting to some 10% of its 150 thousand concepts).

[12] http://alignapi.gforge.inria.fr/

[13] http://aims.fao.org/standards/agrovoc/linked-open-data

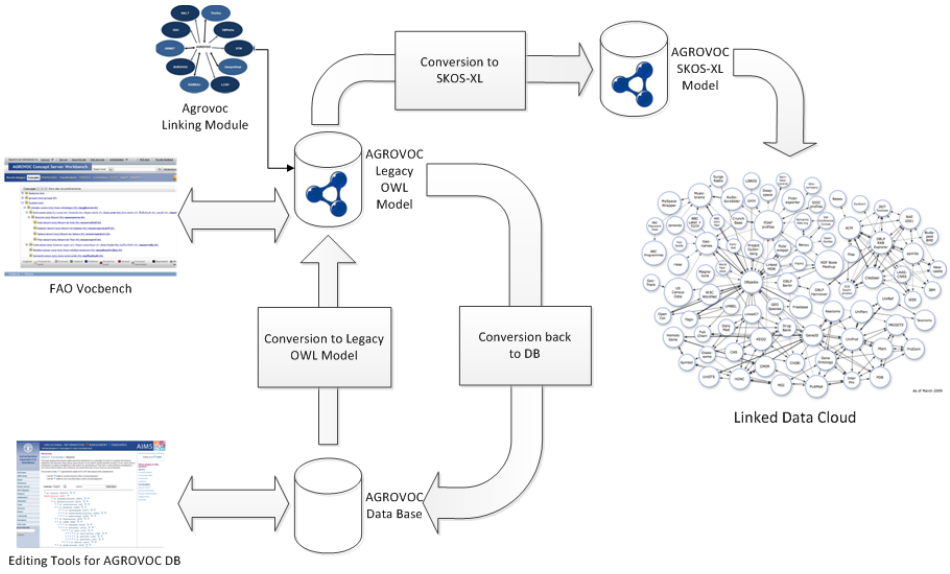

Fig. 5. Overview of the process for publishing AGROVOC as linked data

7 Overview of the Maintenance Process of AGROVOC LD

Fig. 5 provides a high-level view of the entire AGROVOC maintenance process, and its publication as linked data. On the left side of the picture, the tools in use for AGROVOC data maintenance are represented. At the bottom left the legacy web tool based on the relational database is shown, while at the top left there is VocBench. Each allows interaction with a data repository: a relational database (the historical information management system for AGROVOC) and an RDF triple store based on the Protégé API, backed up by the Protégé DB backend in which the data is modeled according to the legacy OWL Model described in section 2.

Note that the relational database is still in use (it serves as a master repository of AGROVOC for many existing applications) and is periodically synchronized with the data repository corresponding to VocBench (see Fig. 5, arrow labeled “conversion back to DB”). However, the data for linked data publication comes from a conversion to SKOS-XL of the data stored in the Protégé DB, according to the legacy OWL model (see Sec. 4).

Given the current situation, publishing AGROVOC as linked data implies a series of steps, many dedicated to data conversion. 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.

Another reason for keeping the relational format and its corresponding applications is that not all editors are able to immediately adopt VocBench. In some cases, this is due to scarce bandwidth, in which case a local copy of VocBench can be used, with batch inclusion in the master copy. In other cases, this is due to the fact that editors continue to use the old tool because they are already well acquainted with it and training efforts for a globally dispersed group of users are complex and resource-intensive.

8 Conclusion

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 specialized tools. In particular, the re-modeling of AGROVOC using OWL and SKOS and it eventual 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 precise tools: editors and workflow managers such as VocBench, triple stores and SPARQL endpoints such as Allegrograph, RDF visualizers such as Pubby, and exotic 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, organizing workshops for data managers, and in improving the functionalities of the VocBench environment so that it can be used by all.

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 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.

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, Sachit Rajbhandari, Prashanta Shrestha, Lavanya Neelam, Jérôme Euzenat, Stefan Jensen, Antoine Isaac, Søren Roug, Thomas Baker, and Mary Redahan.

References

1. Soergel, D., Lauser, B., Liang, A., Fisseha, F., Keizer, J., Katz, S.: Reengineering Thesauri for New Applications: The AGROVOC Example. Journal of Digital Information - JODI 4 (2004)
2. Gennari, J., Musen, M., Fergerson, R., Grosso, W., Crubézy, M., Eriksson, H., Noy, N., Tu, S.: The evolution of Protégé-2000: An environment for knowledge-based systems development. International Journal of Human-Computer Studies 58(1), 89–123 (2003) Protege.
3. Knublauch, H., Fergerson, R., Friedman Noy, N., Musen, M.: The Protégé OWL Plugin: An Open Development Environment for Semantic Web Applications. In : Third International Semantic Web Conference - ISWC 2004, Hiroshima, Japan (2004)
4. W3C: SKOS Simple Knowledge Organization System Reference. In: World Wide Web Consortium (W3C). (Accessed August 18, 2009) Available at: http://www.w3.org/TR/skos-reference/
5. W3C: SKOS Simple Knowledge Organization System eXtension for Labels (SKOS-XL). In: World Wide Web Consortium (W3C). (Accessed August 18, 2009) Available at: http://www.w3.org/TR/skos-reference/skos-xl.html
6. McBride, B.: Jena: Implementing the RDF Model and Syntax Specification. In : Semantic Web Workshop, WWW2001 (2001)
7. Bechhofer, S., Lord, P., Volz, R.: Cooking the Semantic Web with the OWL API. In : 2nd International Semantic Web Conference, ISWC, Sanibel Island, Florida (2003) October.
8. Broekstra, J., Kampman, A., van Harmelen, F.: Sesame: A Generic Architecture for Storing and Querying RDF and RDF Schema. In : The Semantic Web - ISWC 2002: First International Semantic Web Conference, Sardinia, Italy, pp.54-68 (2002) June 9-12.
9. Morshed, A., Caracciolo, C., Gudrun, J., Keizer, J.: Thesaurus alignment for Linked Data publishing. In : Proc. of Dublin Core 2011 (forthcoming) (2011)
10. Cohen, W., Ravikumar, P., Fienberg, S.: A comparison of string distance metrics for name-matching tasks. In : IJCAI-2003 (2003)
11. Euzenat, J.: An API for Ontology Alignment. In McIlraith, S., Plexousakis, D., van Harmelen, F., eds. : The Semantic Web - ISWC 2004: Third International Semantic Web

12 Caterina Caracciolo*, Ahsan Morshed*, Armando Stellato+, Gudrun Johannsen*, Yves Jaques*and Johannes Keizer*

Conference, Hiroshima, Japan, vol. 3298, pp.698-712 (2004) November 7-11.

Additional Readings

1. van Assem, M., Malaisé, V., Miles, A., and Schreiber, G. A Method to Convert Thesauri to SKOS. In The Semantic Web: Research and Applications. 2006, 95-109..
2. Neubert, J.: Bringing the "Thesaurus for Economics" on to the Web of linked data. Proc. WWW Workshop on linked data on the Web (LDOW 2009), Madrid.
3. Zapilko, Benjamin; Sure, York (2009): Converting the TheSoz to SKOS. GESIS Technical Report 2009/07. GESIS-Leibniz Institute for the Social Sciences. ISSN: 1868-905. (2009) URL:http://www.gesis.org/fileadmin/upload/forschung/publikationen/gesis_reihen/gesis_methodenberichte/20
09/TechnicalReport_09_07.pdf.