DCPAPERS
Proc. Int'l Conf. on Dublin Core and Metadata Applications 2015

Interlinking Two Institutional KOS about Agroecology: Using LOD Agrovoc to Circumvent the Language Barrier in Identifying Terminological Intersections

Sophie Aubin
INRA, France
sophie.aubin@versailles.inra.fr

Pascal Aventurier
INRA, France
pascal.aventurier@avignon.inra.fr

Ivo Pierozzi Júnior
Embrapa, Brazil
ivo.pierozzi@embrapa.br

Leandro H. M. Oliveira
Embrapa, Brazil
leandro.oliveira@embrapa.br

Keywords: semantic interoperability; vocabulary alignment; open linked data; metadata; agroecology , skos.

1. Context and Aims of the work

INRA and Embrapa (respectively the French and the Brazilian national institutes for agricultural research) are historical partners in initiatives for knowledge and information management. Given the challenges involved in the mutual sharing of their technical-scientific production especially considering language barriers, efforts have been made to develop semantic interoperability between repositories and bibliographic databases of both institutions. INRA and Embrapa databases (respectively ProdINRA and BDP@) expose bibliographic data with Dublin Core, so the focus of this work was on dc:subject that aims at leveraging by a better semantic interoperability of vocabularies associated with these databases.

Among diverse agricultural subdomains, Agroecology is taking an increasingly important place in the issue of feeding the world, taking into account farmers activity, climate change, and agricultural modernization. Yet, each country and organization has a different understanding of Agroecology and what it covers exactly in terms of social issues, techniques, inputs, and for instance its relation to organic farming. So, considering the ubiquity as well as the ambiguity surrounding the subject, Agroecology was chosen as a case study since both institutions have strategic interests to develop and implement technological facilities to maintain specific terminologies while sharing mutual information. This scenario is extremely timely and demands a quick solution.

This work describes the methodological approach proposed to resolve the matter of indicating equivalent terms in both languages to the same concept recorded in Agrovoc related to the discipline of Agroecology. French and Brazilian vocabularies were not compiled using the same methods and then the analysis was not conducted similarly, requiring different treatment for each vocabulary until the Agrovoc SKOS exact match could be performed.

2. Material and Methods

INRA and Embrapa mutual collaboration aims to share information and knowledge according to their respective technical and methodological aptness. The Semantic Web, with its representation standards and tools, appears to be an interesting meeting point. More specifically, Agrovoc Linked Open Data (Agrovoc LOD) serialized in RDF SKOS and offering concept labels in Portuguese and French was chosen as the key solution. The building of the KOS (Knowledge Organization System) subsets was different for the two institutions.

INRA compiled the list of 3,140 French terms from VocINRA (INRA's own vocabulary) that were used to manually index 2,145 publications about Agroecology in the institutional repository ProdINRA. Onagui (Mazuel and Charlet, 2010), an open source tool designed to help alignment

https://doi.org/10.23106/dcmi.952137170

of vocabularies in SKOS or OWL, was used to align the VocINRA concepts with those in Agrovoc.

Embrapa compiled a Brazilian Portuguese scientific textual corpus from 260 full papers about Agroecology, corresponding to 2,336,287 words and then performed a semi-automatic term extraction and a term matching using a specific tool developed to compare and reuse terminologies and conceptual structures from other KOS (Pierozzi Júnior et al., 2014). From the corpus a preliminary term list was built by semi-automatic term extraction and then it was matched with both Thesagro (a Brazilian Portuguese thesaurus) and Agrovoc-PT, producing a second term list where the exact match terms found in each of the two thesauri were identified and separated and translated in SKOS. The SKOS information from Agrovoc was further retrieved for those terms found at the same time in Thesagro and in Agrovoc.

3. Results

Out of the 3,140 selected French (FR) terms, 1,542 were found in Agrovoc LOD, on the bases of Stoilos (Stoilos et al. 2005) and Levenshtein distance algorithms implemented in Onagui with a nearly exact match distance (0.97). Results using these two string metrics to process the data are probably the same. There is no error in the alignment because it has a chosen value close to the exact match, that is 100% of the words in common. The chosen value for alignment was close to an exact match because the two thesauri are so big that we could not check the alignment values that were too far from the exact match.

Concerning the Brazilian Portuguese (PT/BR) Agroecology vocabulary, the preliminary term list was made up of 783,817 term candidates; the matching with both Agrovoc and Thesagro thesauri resulted respectively in 2,718 and 3,807 terms; exact SKOS match from Agrovoc resulted in 1,699 terms.

The intersection between the FR and PT/BR vocabularies totalizes 939 common URIs from Agrovoc LOD. Some keywords in common are: public health, rural development, recycling, technology transfer, *Raphanus sativus,* root nodulation, soil fertility, sowing depth, tropical climate. Some keywords are specific to INRA research domains as: vineyards, selective grazing, cauliflowers, agrosilvopastoral systems, *Dactylis glomerata*, environmental control. Finally, other keywords are specific to Embrapa: *Araucaria angustifolia*, *Jacaranda*, urban population, molasses, forest inventories, social indicators, passion fruits, root systems, *Leucaena leucocephala*, for example. INRA and Embrapa prepare the results from this work to be publicly available by its webservices.

4. Conclusions and perspectives

The primary motivation for collating INRA and Embrapa methodologies of building and using vocabularies was to implement faster and better semantic interoperability of the KOS used to index and thus share the large amount of scientific knowledge produced in France and Brazil. Agrovoc LOD was chosen to identify and map common terms (and consequently concepts) in the context of French and Brazilian agricultural knowledge using Agroecology as a study case.

Agrovoc LOD proved to be an interesting and feasible solution functioning as a pivot where the methodological differences in the construction of both vocabularies do not interfere in the final result, allowing both (1) the identification of already common terms used for the two institutions and (2) a set of specific terms in each language that might be incorporated into each other's vocabularies. The alignment between the INRA and Embrapa vocabularies prepares these two vocabularies to be linked when published in LOD. Documents contained in the respective institutional repositories of both organizations as well as documents from other institutions around the world, may then be found from the linked data to the Agrovoc URI of a specific term. This work also highlights some difficulties in the translation of certain terms in Agrovoc which will improve this multilingual vocabulary.

https://doi.org/10.23106/dcmi.952137170

This conciliatory methodological model can be strengthened and systematized so that Embrapa and INRA can consider their contribution to broader initiatives in the agricultural domain like the project for a Global Agricultural Concept Scheme (Baker and Suominen, 2014).

References

Agrovoc LOD: http://aims.fao.org/standards/agrovoc/linked-open-data

Baker, Thomas and O. Suominen. Global Agricultural Concept Scheme (GACS): A multilingual thesaurus hub for Linked Data. 2014 http://aims.fao.org/sites/default/files/posts/attachments/GACS_Integration_Proposal_1.0_3.pdf

Mazuel Laurent and Jean Charlet "SPIM-AlignmentGUI - un logiciel d'aide à la réalisation d'alignements entre ontologies 2009. Inria http://ic2009.inria.fr/docs/posters/MazuelCharlet_Poster_IC2009.pdf

Pierozzi, Ivo Júnior, Marcia Izabel Fugisawa Souza, Tércia Zavaglia Torres, Leandro Henrique Mendonça de Oliveira and Leonardo Ribeiro Queiros. Gestão da informação e do conhecimento. In: Tecnologias da informação e comunicação e suas relações com a agricultura. Brasília, DF: Embrapa, 2014. Cap. 12. p. 237-260. URL: http://ainfo.cnptia.embrapa.br/digital/bitstream/item/119627/1/capitulo12-085-14.pdf

OnAGUI - Ontology Alignment GUI :http://sourceforge.net/projects/onagui/

Stoilos Giorgos;Stamou, Giorgos; Kollias, Stefanos (2005). A String Metric for Ontology Alignment. The Semantic Web – ISWC 2005 . Lecture Notes in Computer Science Volume 3729, pp 624-637

https://doi.org/10.23106/dcmi.952137170