View metadata, citation and similar papers at core.ac.uk

brought to you by CORE

provided by Elsevier - Publisher Connector

Available online at www.sciencedirect.com

SciVerse ScienceDirect

Procedia Computer Science 9 (2012) 1571 – 1577

Procedia Computer Science

International Conference on Computational Science, ICCS 2012

Contrasting Knowledge Organization Systems for the description of research products: the case of overlapping in the agricultural domain

Leonardo Lezcano[a*], Elena García-Barriocanal[a], Miguel-Angel Sicilia[a]

[b]*Computer Science Department, University of Alcalá*
Polytechnic Building 28871 Alcalá de Henares, Spain
{leonardo.lezcano, elena.garciab, msicilia}@uah.es

Abstract

The use of Knowledge Organization Systems (KOS) as ontologies or terminologies for the description of scholarly contents requires a careful consideration of the domain and the KOS available. KOS in the same domain may differ in several dimensions including purpose, level of formality, structure and language. In consequence, curators of scientific data face the problem of selecting the relevant KOS, developing mappings when appropriate and deciding on their usage for annotating resources. In domains in which more than a KOS is available, curators need tools to help them in the decision making process. Due to the available heterogeneity of KOS, exploratory tools are required for an initial assessment of overlapping and differences. This paper reports on a practical experience using simple mapping analysis and mapping visualizations in the domain of agriculture. These techniques represent promising directions for the development of decision tools based on the contrast of different KOS metrics.

Keywords: ontology; scholarly resources; AGROVOC; Plant Ontology

1. Introduction

Knowledge Organization Systems (KOS) are schemes for organizing information. According to Hodge (2000), KOS include classification and categorization schemes that organize materials at a general level, subject headings that provide more detailed access, and authority files that control variant versions of key information such as geographic names. KOS also include highly structured vocabularies, such as thesauri, and less traditional schemes, such as semantic networks and ontologies. Different kind of KOS are nowadays used for the description of scholarly resources of diverse kinds (papers, preprints, datasets, etc.), and some KOS are currently exposed in the Web using some form of URI-based identification scheme for concepts or classifiers and in some cases they are accessed through Application Programming Interfaces (API) for using them in diverse applications. It is also expected that many of them progressively move to exposure in the Web of Linked Data (Bizer, Heath and Berners-Lee, 2009),

* Corresponding author. Tel.: +0-000-000-0000 ; fax: +0-000-000-0000 .
E-mail address: leonardo.lezcano@uah.es.

1877-0509 © 2012 Published by Elsevier Ltd. Open access under CC BY-NC-ND license.
doi:10.1016/j.procs.2012.04.172

thus enabling seamless navigation between them via dereferenceable URIs. However, scientific data is often curated in specialized repositories or systems that foster the use of particular KOS via special aids in user interfaces that help users select terms or classes to describe research contents.

Institutional repositories (IR) and other similar kinds of specialized digital collections have become an essential infrastructure for exposing the intellectual inventory of research institutions (Lynch, 2003). In a similar direction, CRIS (Current Research Information Systems) cover the research activity of an organization, following models that expand to a detailed account of organizational structures, researchers, projects and grants among other information entities (Jeffery & Asserson, 2008). Both IR and CRIS typically offer search and browsing functionalities to their users, some of them mediated by terminologies and/or ontologies. In consequence, the decision of using some particular KOS has an impact on the organization and functionality of scholarly content curation systems, and thus it requires a previous analysis of the options available. The problem of deciding which KOS is better fitting a particular collection involves the study of their degree of overlapping and extent to which the combination of KOS is adding real value to the solution.

In other cases, an approach considering a single KOS would be a better option, and interoperation can be achieved through KOS-to-KOS mappings. In any case, curators face the challenge of taking those decisions without specific tools. Expertise in the history, coverage and other aspects of each KOS is required for the task, and in many cases these are not available. An option for helping curators would be that of developing analysis tools that could be used to have a quick impression on the similarities of the different KOS as a point of departure for a more comprehensive evaluation. The simpler way of contrasting two KOS is by examining their mappings at the lexical level. As lexical matching is usually covering a large amount of actual conceptual overlapping, it has the merit of giving a first impression of the degree of complement or overlapping between different KOS. The mappings can then be used for inspecting particular approaches to modelling, as well as to generate visual representations showing overlapped areas in KOS hierarchies.

This paper reports on a concrete approach for the task of evaluating KOS overlapping and coverage in the domain of agriculture. While the results are domain-specific, applied techniques can be further used to develop tools that ease KOS evaluation by tailoring them to the needs of scientific data curators. The rest of this paper is structured as follows. Section 2 provides background information on the ontologies and terminologies used in the presented case. Then, Section 3 describes the mapping and analysis of overlapping between the previously introduced KOS. Additional analysis using visualizations is provided in section 4. Finally, conclusions and outlook are provided in Section 5.

2. Background

AGROVOC is a comprehensive multilingual agricultural thesaurus. Organized as a concept scheme, it contains almost 40,000 concepts in over 20 languages covering subject fields in agriculture, forestry and fisheries, together with crosscutting themes such as land use, rural livelihoods and food security. Concept schemes are more flexible than traditional vocabulary models. They are able to handle taxonomies, controlled vocabularies, and subject headers. AGROVOC standardizes data description to enable a set of core integration goals: interoperability, reusability and cooperation. It is maintained by a global community of librarians, terminologists, information managers and software developers.

The AGROVOC concept scheme can be represented by means of the Simple Knowledge Organization System (SKOS[b]) formal language, which is commonly used and can be interpreted by a wide variety of existing systems.

There are three levels of representation:

- *concepts* (the abstract meaning), for example “maize” in the sense of a cereal;
- *terms* (language-specific lexical forms), for example “maize”, “maíz” or “corn” in order to provide a multilingual support;

[b] http://www.w3.org/TR/skos-reference/

- *term variants*, for example "organization" or "organisation", "cow" or "cows", "Zea mays" or "Z. mays", thus providing a range of forms that can occur for each term such as spelling variants, singular, plural, abbreviations, etc.

The abstract concepts build the actual structure of the concept scheme that is represented by all the terms in all languages to which the concept scheme is associated. The entire representation of a concept often includes many terms and both concepts and terms participate in relationships with other concepts and terms. AGROVOC contains two types of relationships:

- Inter-level relationships, which are in turn divided in:
  - Concept-to-Term relationships such as *has lexicalization* that link concepts to their lexical realizations, e.g., the "rice" concept links to its language-dependent terms such as "rice", "riz", "arroz", "paddy", i.e., *(concept) rice has lexicalization (English term) rice.*
  - Term-to-String relationships such as *has acronym*, *has spelling variant* and *has abbreviation* that link language-dependent terms to their variants, e.g., *African Union has acronym AU*.
- Intra-level relationships, which are in turn divided in:
  - Concept-to-Concept relationships such as *has subconcept* and *is used to make* that relate two different ideas, e.g., *cereals has subconcept maize*, *maize is used to make corn flour*.
  - Term-to-Term relationships such as *has synonym* and *has scientific name* that relate two terms that belong to the same concept, e.g., *maize has synonym corn* and *beetles has scientific name Coleoptera*.

AGROVOC has been recently published as Linked Open Data (LOD) in order to connect the different knowledge organization systems in the agricultural domain. Now it has 21,000 outlinks and 3,000 inlinks according to FAO's AGROVOC website. The new resources to which AGROVOC is linked include: EUROVOC, NALT, GEMET, RAMEAU, LCSH, STW, TheSoz, DBpedia, Geopolitical Ontology and DDC.

The Plant Ontology (PO) allows users to ascribe attributes of plant structure (anatomy and morphology) and developmental stages to data types, such as genes and phenotypes, to provide a semantic framework to make meaningful cross-species and database comparisons (Avraham et al., 2008). The Plant Ontology Consortium (POC) builds upon previous work by the Gene Ontology Consortium (GOC) by adopting and extending the GOC's principles, existing software and database structure. PO is currently based on a small subset of concepts and relations. Relations are specific to the particular domains addressed. The main relations are the following:

- **is_a:** The *is_a* relation is used to indicate the relationship between a specific class and a more general one. For example, *megasprophyll is_a sporophyll* and *sporophyll is_a phyllome*.
- **part of:** The *part_of* relation is used to indicate that one class is part of another class. For example, *ectocarp is part_of pericarp*, which in turn *is part_of fruit*.
- **derives_from:** The *derives_from* relation is used to indicate that one plant structure succeeds another across a temporal divide in such a way that at least a biologically significant portion of the matter of the earlier structure is inherited by the later. For example, the fact that *leaf-derived cultured plant cell derives_from leaf* indicates that a significant portion of the matter of a *leaf-derived cultured plant cell* is inherited from some cell in a *leaf*.
- **develops_from:** The *develops_from* relation is used to indicate that a plant structure develops from its parent term. For example, *root hair cell develops_from trichoblast.*
- **adjacent_to:** The *adjacent_to* relation is used when one plant structure is in permanent contact with another one. For example, *anther wall endothecium adjacent_to anther wall exothecium*. In this example, every instance of *anther wall endothecium* should be in permanent contact with (*adjacent_to*) some instance of *anther wall exothecium*. This does not imply that every *anther wall exothecium* is *adjacent to* some *anther wall endothecium*. If the latter were also true, that relation would have to be asserted separately. The *adjacent_to* relation is not transitive.
- **participates_in:** The *participates_in* relation is used to indicate that an anatomical entity only occurs during a particular plant growth or development stage. For example *archegonium participates_in gametophytic phase* and *vascular tissue participates_in sporophytic phase.*

The Organic.Edunet (OE) ontology is the ontology developed to support the Organic.Edunet portal[c]. Within this portal, the Organic.Edunet ontology is mainly used for two purposes:

- Resource annotation: each time a content provider insert a resource in the repository, the resource is annotated with one or more concepts extracted from the ontology.
- Resource retrieval: when web users perform queries on the system, the ontology is used in order to perform advanced searches based on semantic techniques.

The OE ontology is structured in four different sub-ontologies that cover different aspects of the Organic Agriculture domain. They are:

- Issues: they represent all matter of discussion, debate, or public concern in the agricultural domain.
- Activities: they represent the list of abstract concepts used to serve as the basis for more specific concepts in the OE ontology.
- Methods and techniques: they represent the list of concepts giving support to most of the methods and practices used in agriculture.
- Products: they represent the result of agricultural activities. These concepts are split into three sub-concepts: processed products (they are the results of one or more processes of manipulation, e.g. beer, cheese, etc.), unprocessed products (all kinds of vegetables, chemical compounds and products of an animal origin), and fertilizers.

3. Studying overlapping: insights from the VOA3R project

The principal use of ontologies and terminologies, as described above, is that of annotating the resources inside repositories of scholarly content, and then exploiting those annotations for purposes of information filtering in general (search, navigation, recommending resources, etc.). The following reports the analysis carried out by lexical mapping of PO, AGROVOC and the OE ontology. In addition to concepts, the OE ontology also defines a set of relationships based, and subsequently adapted, on the set of basic relationships described in the FAO AGROVOC thesaurus, e.g. *actsUpon*, *affects*, *afflicts*, etc.

The OE ontology is targeted to one of the user communities defined in the VOA3R project. In order to define its usefulness in the new context where AGROVOC was used as a foundation, an inquiry was carried out to identify the extent to which AGROVOC could be used as a replacement of this ontology. Evidence was found in two issues: (a) it is difficult to determine the frontiers between terms that are specific to organic agriculture and those that are not, and (b) the OE ontology contains terms that are not exclusive to organic agriculture and agroecology.

AGROVOC terms	OE ontology terms	Exact matches	Partial matches (AGROVOC term as substring)	Partial matches (OEO term as substring)
40,905	289	98 (34%)	262	1,303

Table 1. OE ontology – AGROVOC lexical mapping results.

A formal approach to studying the overlapping of the OE ontology and AGROVOC was conducted. Table 1 shows the main figures of a process of automated mapping. It should be noted that the OE ontology labels that were formed by the concatenation of two or more words (e.g. LocalBreeds) were split into a phrase, by inserting a space between words (i.e. Local Breeds), in order to match the AGROVOC terms, as they already contain spaces. It is especially relevant that about 34% of the OE ontology terms (classes and instances) have an exact match in AGROVOC. The inspection of the correct mappings found revealed that they occur at different levels of the

[c] http://portal.organic-edunet.eu/

hierarchy of AGROVOC, so they do not appear to be associated to some particular sub-tree(s), indicating that there is a high degree of overlapping in general terms.

These findings led to the decision of not formally including the OE ontology in the VOA3R service. Further, the OE ontology is in plans to be moved to a linked data approach based on the Moki tool[d] (explicitly including a formal mapping to AGROVOC), which will make it homogeneous with the above mentioned approach and consistent with VOA3R principles. Even though PO has a very specific and narrow focus, evidence of overlapping with AGROVOC was found. A formal approach to studying the overlapping of the Plant Ontology and AGROVOC was conducted. Table 2 shows the main figures of a process of automated mapping.

AGROVOC terms	PO terms	Exact matches	Partial matches (from AGROVOC terms)	Partial matches (from PO terms)
40,905	1,450	49	400	630

Table 2. PO – AGROVOC lexical mapping results

In spite of the fact that matches are not in a significant amount, it is worth analyzing if PO could be considered an extension of some of AGROVOC sub-trees, in which case mapping could be an option. For doing so, graphical representations were used as described in Section 4.

4. Visualizing overlapping

For the case of AGROVOC-PO mapping, a graphical visualization was used to illustrate some of the discovered patterns. From the 40950 terms in AGROVOC, near 400 were mapped to PO, but given that all AGROVOC terms cannot be rendered in the same image without losing details, Figure 1 only shows 2265 nodes representing the descendants of the AGROVOC “ENTITIES” term. From the 154300 relations included in AGROVOC (broader, narrower, related, etc), only *broader* and *narrower* are shown for the "ENTITIES" descendants, i.e. 2264 edges. Mappings to PO are represented in Figure 1 as black nodes. It should be noted that the bigger the black node, the more PO terms have been mapped to the AGROVOC term represented by that node.

Near 50% of PO terms were mapped to AGROVOC while only 0,9% of AGROVOC terms were mapped to the PO, proving that the PO covers a narrow and more specialized domain than a comprehensive agricultural resource as AGROVOC. The fact that the AGROVOC terms that were mapped to PO are organized in small and dense areas within the AGROVOC radial tree visualization corroborates such idea. In spite of such conclusion, those areas of high density of terms mapped to PO include a few unmapped AGROVOC terms whose study may lead to one of two results: either they can be mapped to PO or their inclusion as new PO terms should be considered.

In addition to the high density, AGROVOC terms mapped to PO are, in most cases, vertically organized in thin columns that cover most the descendants of a given AGROVOC term in the AGROVOC radial tree visualization. The hierarchy has been built based on the NARROWER and BROADER relations of the AGROVOC ontology which are semantically similar to the popular IS-A relationship. Such vertical concurrence shows that, in spite of covering a very small fraction of AGROVOC, the levels of abstraction of PO are similar to the ones of AGROVOC.

The PO comprises terms covering two biological domains: i) terms that describe morphological/anatomical structures of plants; and ii) terms that describe stages in the growth and development of an entire plant. A result of the current research is the fact that 77% of the PO terms mapped to AGROVOC belong to the Plant Anatomy ontology (one of the two PO sub-ontologies) while only 23% of them belong to the Plant Development ontology (the other PO sub-ontology). Such distribution reveals that the AGROVOC-PO overlapping is much more common in the structure domain than in the process domain.

[d] https://moki.fbk.eu/website/index.php

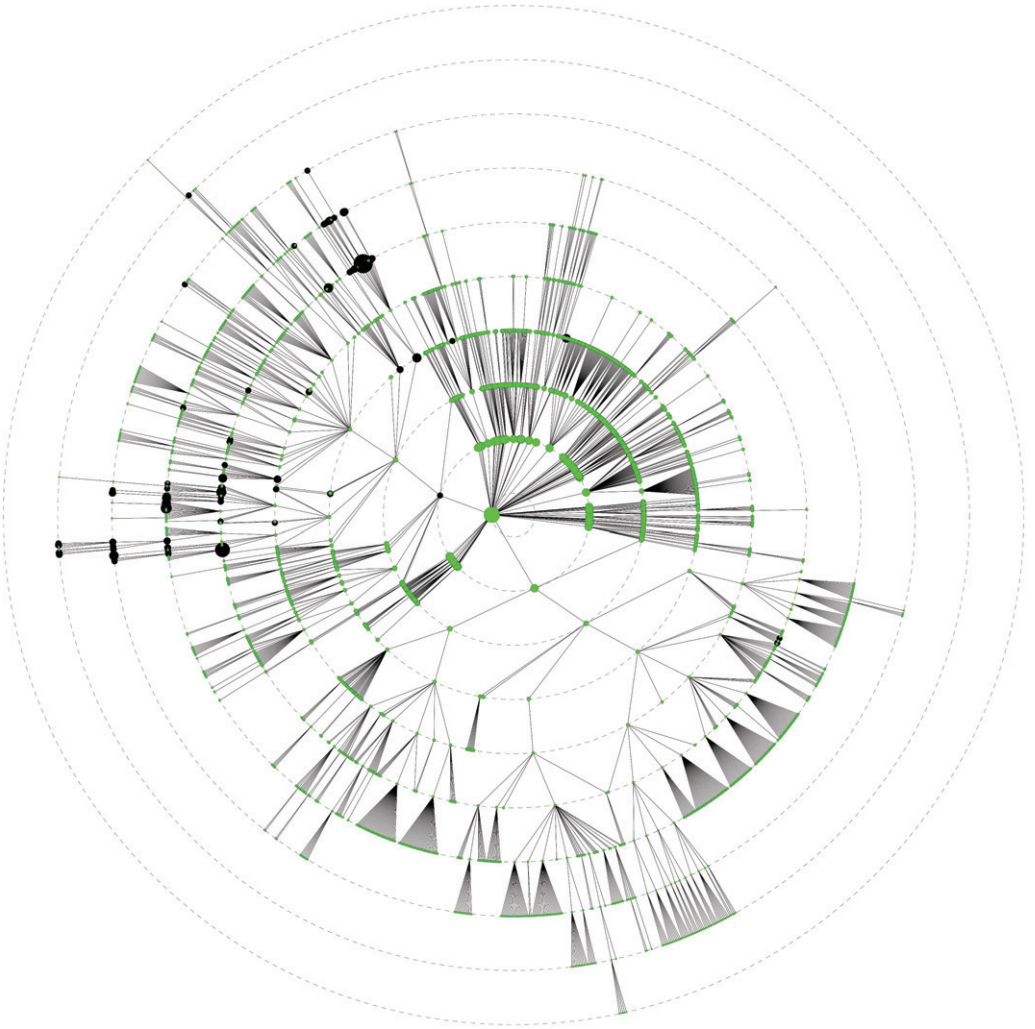

Figure 1. Simplified visualization of the position of mappings of PO to AGROVOC

5. Conclusions and outlook

Knowledge Organization Systems are nowadays diverse and heterogeneous, and data curators can expect finding different KOS in the same domain with a degree of overlapping. This calls for tools and techniques that help curators in assessing the degree of complementarity of several KOS. This paper has reported a case study in the domain of agriculture showing the benefits of two very simple assessment techniques: lexical mapping and its visualization. The use of lexical mappings was able to identify large degrees of overlapping in KOS, and visualization provided useful hints in identifying areas of complementarity.

Future work should be developed in further testing these simple techniques and refining them into tools for curators, mixing with ontology/terminology metrics and looking for heuristics that are able to inform overlap and complementarity for particular sub-domains.

Acknowledgements

The work leading to these results has received funding from the European Commission under grant agreement nº 250525 corresponding to project VOA3R (Virtual Open Access Agriculture & Aquaculture Repository: Sharing Scientific and Scholarly Research related to Agriculture, Food, and Environment), http://voa3r.eu.

References

1. Avraham, S. et al. (2008) The Plant Ontology Database: a community resource for plant structure and developmental stages controlled vocabulary and annotations Nucl. Acids Res. 36 (suppl 1): D449-D454.
2. Bizer, C., Heath, T. and Berners-Lee, T.: Linked data - the story so far. International Journal on Semantic Web and Information Systems (IJSWIS), 5(3): 1-22 (2009).
3. Hodge, G. (2000). Systems of Knowledge Organization for Digital Libraries. Beyond traditional authority files. Washington, DC: the Council on Library and Information Resources
4. Jeffery, K., & Asserson, A. (2008). Institutional Repositories and Current Research Information Systems. New Review of Information Networking, 14(2), 71-83.
5. Lynch, C. (2003). Institutional Repositories: Essential Infrastructure for Scholarship in the Digital Age. *Portal: Libraries & the Academy*, **3(**2), pp. 327–336.