IFLA WLIC 2017 WROCŁAW

Submitted on: 03.07.2017

Optimizing subject access to legal resources: EuroVoc and VocBench. Benefits of using multilingual controlled vocabularies and an open source collaborative tool for their maintenance

Rudolf W. Strohmeier
Publications Office of the European Union, Luxembourg

Willem van Gemert
Publications Office of the European Union, Luxembourg

Christine Laaboudi
Publications Office of the European Union, Luxembourg

Anne Waniart
Publications Office of the European Union, Luxembourg

The opinions expressed in this document are those of the authors and do not represent the views of the European Commission nor of any other organisation.

Abstract:

The Publications Office of the European Union is an interinstitutional office entrusted with the task of publishing EU law and publications in the 24 official EU languages. To face the challenge of publishing multilingual legal data, the Publications Office makes use of multilingual controlled vocabularies such as authority tables and the EuroVoc thesaurus.

EuroVoc is a multilingual multidisciplinary thesaurus covering the activity fields of the EU institutions. It is expressed using the semantic web standards recommended by the W3C organization. Besides modelling flexibility and reasoning capabilities, these technologies also facilitate the alignment of EuroVoc with other thesauri and controlled vocabularies, such as national and domain-specific vocabularies.

Authoring, maintenance and management of the EuroVoc thesaurus is performed using VocBench, an open source collaborative web platform for the development of thesauri, complying with the SKOS and SKOS-XL standards. Thanks to funding by the ISA^2 programme of the European Commission a new major version of VocBench has been developed that will be available in Q3/2017.

Keywords: controlled vocabularies, thesaurus, subject access, semantic web standards

About the Publications Office of the European Union

The Publications Office of the European Union is an interinstitutional office set up by EU institutions in 1969 and entrusted with the task of publishing EU law and EU publications, in all 24 official languages of the European Union. By doing so, it contributes to a practical implementation of the European Union's motto "United in diversity", which signifies how Europeans have come together to work for peace and prosperity, while at the same time being enriched by the Europe's many different cultures, traditions and languages.

In this context, as a modern multilingual information provider, the Publications Office offers state-of-the-art information management services to all EU institutions, agencies and bodies, as well as to EU citizens at large. Having established itself as a recognized standardization actor, it has also taken a lead role in standardizing metadata and formats for the exchange of information, including legal data. The Publications Office also administers a number of information management tools that are used to this end by the EU institutions, agencies and bodies, national and regional administrations in Europe, as well as national governments and private users around the world.

A substantial part of the information published by the Publications Office is available in the 24 official languages of the European Union, for example all legal acts published on EUR-Lex[1], the gateway to EU law, and part of the EU publications, disseminated online via EU Bookshop[2].The multilingual aspect of the production of the Publications Office poses a number of challenges in regard to optimizing subject access. One of the solutions implemented by the Publications Office to address these challenges is the use of multilingual controlled vocabularies with language-independent identifiers. These vocabularies can be found in the Publications Office Metadata Registry[3]. All of these tools are provided as open data, which can be reused by any interested parties in Member States of the European Union and beyond. By providing such tools the Publications Office facilitates mutual understanding and interoperability between divergent legal systems, thus contributing to the creation of a common European space for collaboration and mutual learning in this area.

EuroVoc

EuroVoc is the most useful controlled vocabulary for optimizing access to subject matter in EU and national legal data, as well as for other information sources. It is a multidisciplinary thesaurus that offers around 8000 concepts in 23 official EU languages, plus three more, corresponding to EU candidate countries. It covers all activity fields of the European Union institutions. The concepts are classified according to EU policies and fields of activity in 21 domains and 120 microthesauri (sub-domains). EuroVoc serves two content-indexing purposes: EU legislation, in EUR-Lex; and EU publications, disseminated by the Publications Office via EU Bookshop and OPac[4] (the Publications Office online public access catalogue). Since EuroVoc concepts are language-independent, users can search for a concept in their own language and get the results in a different language.

[1] http://eur-lex.europa.eu/.

[2] http://publications.europa.eu/bookshop.

[3] http://publications.europa.eu/mdr/.

[4] http://opac.publications.europa.eu/search/query?theme=system.

EuroVoc is compiled in accordance with the standards of the International Standards Organisation ISO 25964 – *Thesauri and Interoperability with other Vocabularies*. It is continually adapted to keep up with three changing paradigms: the fields in which the EU institutions are concerned; the language arrangements; and the technological innovations. The maintenance team at the Publications Office collects and examines the proposals coming from the national parliaments, the European institutions or agencies, and private users (citizens or companies). A governance process has been established to deal with the administration and maintenance of the thesaurus, with updates to EuroVoc being published, in principle, biannually.

EuroVoc is accessible for browsing and searching through its website (http://eurovoc.europa.eu). It is also available for download from the EU Open Data Portal (https://open-data.europa.eu/en/data/dataset/eurovoc) in machine-readable formats such as XML and SKOS/RDF. Additionally, a number of web services allow users to query EuroVoc and to display or use the results directly in their applications: http://eurovoc.europa.eu/drupal/?q=webservice&cl=en. The Publications Office is a pioneer in using semantic technologies in its production and dissemination chains. EuroVoc is available through the SPARQL endpoint of the Publications Office and can be used together with other legal resources made available by the Publications Office for reuse purposes and linking data.

EuroVoc, which has become the reference tool for indexing EU information, is used by EU institutions, agencies and bodies, national and regional parliaments and governments in Europe, public administrations, libraries, information professionals and private users around the world.

Towards the semantic web

Linked Open Data (LOD) is increasingly becoming a de-facto standard and a set of practices in the data publishing world. This is mainly thanks to a rich set of standards (RDF-S, OWL, etc.); widely accepted ontologies (Dublin Core, DCAT, SKOS(-XL), schema.org, etc.); and a strongly involved community.

EuroVoc is one of the first multilingual vocabularies to make use of semantic web technologies that reflect W3C recommendations and the latest trends in thesaurus standards. Originally EuroVoc was based on the SKOS (Simple Knowledge Organization System) and SKOS-XL (SKOS extension for Labels) ontologies. Gradually the content information evolved beyond the simple SKOS and SKOS-XL models towards a new model representing the Publication Office business needs. In order to enable interoperability and to promote wide usability of published data, the Publications Office has developed an application profile (SKOS-AP-EU), to shape thesauri and controlled authority lists[5]. The SKOS-AP-EU application profile extends SKOS with properties from a range of well-known vocabularies. This application profile is formally expressed using SHACL (shape constraint language). It allows for the automatic validation of controlled vocabularies and the simultaneous generation of documentation.

[5] Source: Costetchi E., Van Gemert W. (2016).

EuroVoc as linked data

Mapping EuroVoc to more specialised controlled vocabularies

The fact that EuroVoc is available in semantic formats such as SKOS provides an excellent opportunity to map and align it with more specialised vocabularies.

EuroVoc covers a number of different disciplines at a generic level and therefore cannot always reflect the specificity required by experts in a domain. To ensure broader coverage, EuroVoc is aligned with more specialised thesauri, such as the thesaurus of the Food and Agriculture Organisation (AGROVOC), the thesauri of the United Nations (UNBIS and UNESCO), the STW Thesaurus for Economics maintained by the Leibniz Information Centre for Economics, etc. The Leibnitz Foundation is going to integrate the EuroVoc labels mapped with the STW thesaurus in order to enrich their search engine.

GEMET, the GEneral Multilingual Environmental Thesaurus, has been developed as an indexing tool for the European Environment Agency (EEA)[6]. The semantic mapping of GEMET with EuroVoc will provide added value by facilitating access to documents in the environmental domain, which is a key area of EU activities. Both thesauri are multilingual (23 languages for EuroVoc, and 35 for GEMET), with different levels of abstraction and detail. For example: GEMET uses the concept “environmental legislation”[7] and the semantic correspondence in EuroVoc is “environmental law”[8].Thanks to the mapping of the two thesauri, users could combine familiar terminology from one of the two thesauri in their native languages and let the system provide the correspondences to other terminologies in order to retrieve equivalent results.

Mapping EuroVoc with national controlled vocabularies to annotate national legal information

EuroVoc covers a number of different disciplines at a generic level and, therefore, cannot reflect the differences between distinct legal systems at a sufficiently detailed level. However, linking EuroVoc and the controlled vocabularies of legal systems can expand the generic legal concepts, or the EU concepts from EuroVoc, into the terms used in the national legislation of a given country. As a result, both vocabularies can be used to optimize subject matter access in either the national legislation or the EU legislation (via EUR-Lex). Another advantage of thesauri alignment is the extension of the search strategy to other national or European legal databases.

The benefit of the alignment of a national vocabulary with EuroVoc can be summarized as follows: if a concept of a legal system has an equivalent in EuroVoc, it is automatically translated; if the national concept is too specific, it can be mapped with a more generic EuroVoc concept. For example: the management of the Moselle River concerns three countries (France, Germany and Luxembourg). In Legilux, Luxembourg's official online-law website, the concept to be used is “canalisation de la Moselle”[9]. EuroVoc describes the geography of the Member States only at regional level and does have equivalence for this very specific concept. By aligning the specific term "canalisation of the Moselle” with the

[6] https://europa.eu/european-union/about-eu/agencies/eea_en.
[7] http://www.eionet.europa.eu/gemet/concept?ns=1&cp=2862.
[8] http://eurovoc.europa.eu/535.
[9] http://legilux.public.lu/search/A/?fulltext=&thematique=canalisation+de+la+moselle.

existing EuroVoc concept "inland waterway", the search strategy could be extended to enhance the comparison of national and/or European legal data.

Official collaborations exist amongst Member States and the European Union for the use of EuroVoc to index national government and parliamentary data. For example, the National Council of the Slovak Republic [10] has implemented Eurovoc in the context of its parliamentary information system. The National Assembly of the Republic of Slovenia[11] uses EuroVoc as a basis for the code list of the government's EU portal. The Belgian Parliament and Senate[12] use Eurovoc as controlled vocabulary to search in their respective databases.

Another example of this type of collaboration, undertaken in the context of the European Legislation Identifier (ELI), is a project led by the Official Journal of the Grand Duchy of Luxembourg and the Publications Office. This initiative has aligned the concepts in EuroVoc with the controlled vocabulary used to index the content of the national legislation of the Grand Duchy of Luxembourg published on http://legilux.public.lu/. The purpose of the project was to make the EU and the national legislation of Luxembourg interoperable, and to facilitate access to legal resources. The results confirm that using EuroVoc as a bridge between more specialised controlled vocabularies of national legal systems can bring distinct advantages in optimizing access to legislation.

VocBench

The Publications Office uses *VocBench* for the maintenance of EuroVoc. *VocBench* is a web-based open-source collaborative platform for the management of multilingual controlled vocabularies which makes use of semantic technologies and complies with the SKOS and SKOS-XL standards. It is particularly suitable for managing large thesauri in RDF format.

The Publication Office is an active member of the large multilingual thesauri community and drives forward the development of *VocBench* to the next version. The latest major version of *VocBench 3.0*, developed thanks to funding by the ISA^2 programme of the European Commission, contains a number of new functionalities, including built-in alignment features. Soon available in Q3/2017, it will offer users the necessary toolbox to create, maintain, link and publish their controlled vocabularies, metadata or glossaries as Linked Open Data. *VocBench 3.0* will offer a general purpose collaborative environment for development of any kind of RDF dataset; improving the editing capabilities of its predecessor, while still maintaining the peculiar aspects that determined its success.

VocBench is open source and can be used by any public organisation, company or independent user to manage their controlled vocabularies. The latest installation package as well as building and deployment instructions for *VocBench* 2.x/3.0 can be found directly on the project site.

[10] http://kniznica.nrsr.sk/index.php/en/eurovoc.

[11] http://www.mzz.gov.si/en/foreign_policy_and_international_law/eu_policies/language_issues/legal_and_linguistic_databases/eurovoc/.

[12] http://www.dekamer.be/kvvcr/showpage.cfm?section=/flwb&language=nl&cfm=ListKeyword.cfm?legislat=54 and https://www.senate.be/www/?MIval=/index_senate&MENUID=12000&LANG=fr.

Conclusion

Ensuring optimized subject access to legal data from a wide range of divergent sources, such as EU institutions and the publishers of national legislation from EU Member States, is an essential component of the multilingual legal information ecosystem in the European Union, which thrives on fostering unity among its diverse members. Thanks to using open standards and making the relevant tools presented in this article available for reuse worldwide, the Publications Office goes further and promotes the idea that embracing solidarity in divergence is a valid and valuable proposition which can bring substantial benefits to those who are ready to put it into practice.

References

Costetchi E., Van Gemert W. (2016). Towards executable application profile for European vocabularies. W3C. https://www.w3.org/2016/11/sdsvoc/SDSVoc16_paper_23.

Publications Office of the European Union (2010). Eurovoc Conference: Mind the Lexical Gap. EU. Publications Office. http://eurovoc.europa.eu/drupal/?q=node/936.

Stellato A. et al. (2015). VocBench: A Web Application for Collaborative Development of Multilingual Thesauri. The Semantic Web. Latest Advances and New Domains. ESWC 2015. Lecture Notes in Computer Science, vol 9088. Springer, Cham. http://link.springer.com/chapter/10.1007/978-3-319-25639-9_29.