

Aligning Legivoc Legal Vocabularies by Crowdsourcing

Hughes-Jehan Vibert, Benoît Pin, Pierre Jouvelot

▶ **To cite this version:**

Hughes-Jehan Vibert, Benoît Pin, Pierre Jouvelot. Aligning Legivoc Legal Vocabularies by Crowdsourcing. Language and Semantics Technology for Legal Domain (LST4LD) Workshop, 10th Recent Advances in Natural Processing Conference, Sep 2015, Hissar, Bulgaria. hal-01251086

HAL Id: hal-01251086

https://minesparis-psl.hal.science/hal-01251086

Submitted on 5 Jan 2016

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.

Aligning *legivoc* Legal Vocabularies by Crowdsourcing

Hughes-Jehan Vibert

Ministère de la Justice, France
hughes-jehan.vibert@justice.gouv.fr

Benoit Pin, Pierre Jouvelot

MINES ParisTech, PSL Research University, France
{benoit.pin, pierre.jouvelot}@mines-paristech.fr

Abstract. *legivoc* is the first Internet-based platform dedicated to the diffusion, edition and alignment of legal vocabularies across countries. Funded in part by the European Commission and administered by the French Ministry of Justice, *legivoc* offers a seamless path for governments to disseminate their legal foundations and specify semantic bridges between them. We describe the general principles behind the *legivoc* framework and provide some ideas about its implementation, with a particular focus on the state-of-the-art tools it includes to help crowdsource the alignment of legal corpora together.

1 Introduction

legivoc[1] (all in lower case) is an Internet-based database platform dedicated to the management of multiple legal information terminologies, with a particular focus on vocabularies and their alignments [4]. The system is designed to be used both interactively and also as an automated Web service, interoperable with other document management tools or international legislation or translation systems, via a dedicated Application Programing Interface (API).

The main goals of *legivoc* are: (1) to provide access, within a unique framework and using a general formalism, to (ultimately) all the legal vocabularies of the Member States of the European Union; (2) to foster the use of best practices regarding the encoding of these vocabularies using Internet standards such as the Simple Knowledge Organization System (SKOS) and Uniform Resource Identifier (URI); (3) to encourage the creation of alignment information between these vocabularies, helping provide bridges between judicial systems based on different laws and languages.

The French Ministry of Justice spearheads the project, partly funded by the European Commission and the Ministries of Justice of the Czech Republic, Spain, Finland, France, Italy and Luxembourg. ARMINES and MINES ParisTech are the lead scientific advisors and implementation specialists for the *legivoc* project.

[1] `http://legivoc.org` (the site is open, although in an "alpha" version).

legivoc is intended to be used directly by law and judicial experts, for instance when dealing with cross-border legal issues or planning new legislative regulations. For such a purpose, knowing how a legal notion in a given vocabulary relates to similar ones in other countries, a so-called alignment is a key asset. We show how *legivoc* can be extended so that entering such information is a very intuitive operation, with the idea of relying on crowdsourcing efforts (where dedicated individuals perform useful tasks for the community, as in Wikipedia) to enrich its vocabularies.

The rest of this paper is structured as follows. In Section 2, we provide a brief introduction to *legivoc* structure and typical use cases. Section 3 focuses on the alignment process, which intends to build upon international crowdsourcing efforts to enrich the *legivoc* database. Section 4 outlines the communication capabilities embedded within *legivoc*, for use by the systems that want to take advantage of it. We conclude in Section 5.

2 *legivoc* 1.01

legivoc materializes as a Web site providing a state-of-the-art multilingual system for the creation, edition and alignment of international legal vocabularies (see Figure 1 for the site page).

2.1 User Interface and Capabilities

In its current (6/2015) state, *legivoc* enables registered users to access the legal vocabularies of 13 Member States plus Switzerland, in addition to a global one managed by Eurovoc[2]. These vocabularies exist in multiple languages; these translated versions have been either provided by the Member States themselves or automatically translated via the European Commission MT@EC multilingual service [3].

Words in vocabularies are considered as SKOS concepts. They can be (1) visualized in various forms (text, dendogram, SKOS source), (2) edited, (3) related to more or less abstract concepts or (4) aligned to similar concepts in other vocabularies (see Section 3). Figure 2 illustrates part of a typical display of a concept, here the one of "civil law"; it was obtained, after a search for "droit civil" in *legivoc*, via the French version of the Eurovoc vocabulary, identified by its International Standard Organization (ISO) code `eu`, followed by its *legivoc* number, 523.

[2] `http://eurovoc.europa.eu`



Fig. 1. *legivoc* home page

2.2 Technological Foundations

legivoc is built on top of plinn[3], an open-source collaborative infrastructure developed at ARMINES and MINES ParisTech. This environment, based on the Zope multi-purpose Web framework system, provides sophisticated ways to manage documents in various formats, with a particular emphasis on users' access rights, for security purposes, and workflows, to enforce best practices. The existing legal vocabularies provided by the State Members use various formats (ASCII, Excel, Word, XML, SKOS…); thus heuristics have been developed and implemented to encode them into a common *legivoc*-specific SKOS format. Introducing new legislative corpora, e.g., from countries wishing to join *legivoc*, requires the design and implementation of new approximate algorithms, unless the data is already encoded into the *legivoc* SKOS format, which is strongly advised.

Even though the presence of such diverse formats and possibly incompatible semantics is a clear challenge and the existence of a general, widely adopted vocabulary format would be a desirable feature, we take, with *legivoc*, a "can-do" approach, and try to leverage existing corpora instead of waiting for an eventual standard. In fact, our own *legivoc* SKOS format can be seen as a first proposal in this direction.

[3] `http://www.plinn.org`

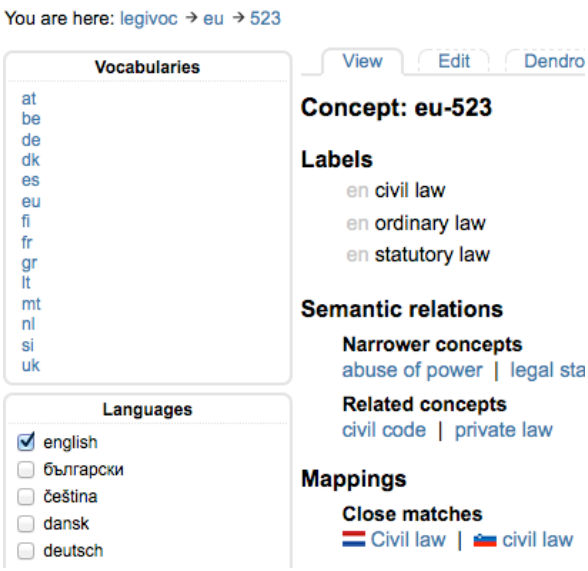


Fig. 2. Concept for "droit civil" (excerpt)

legivoc strongly adheres to W3C standards, such as (1) the Resource Description Framework (RDF), which uses 3-tuples (subject, predicate, object), also called triples, to represent data, (2) the URI naming convention, to denote such resources, (3) SKOS, an RDF-based representation format for vocabularies, and (4) the SPARQL Protocol and RDF Query Language (SPARQL), to search and access triples in an effective and expressive manner.

Figure 3 shows how IT specialists can write and execute (possibly from a remote site, see Section 4) *legivoc* SPARQL commands to answer specific requests emanating from law or judicial experts.



Fig. 3. Querying the Greek vocabulary

In this particular example, one asks for the list of all concepts, given by their URI, and their preferred labels, as strings, present in the Greek vocabulary. An excerpt of the output is given in Figure 4.

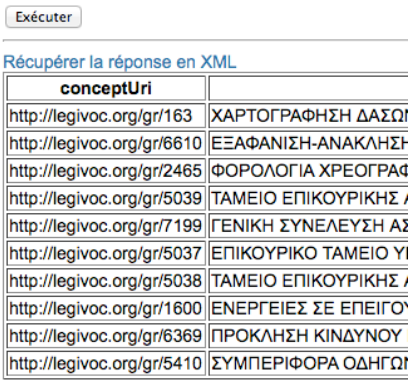


Fig. 4. Greek concepts and labels (excerpt)

3 Alignment Management

Alignment is a feature supported by SKOS, via the predefined `closeMatch` (see an example in Figure 2), `exactMatch`, `broadMatch`, `narrowMatch` and `relatedMatch` mapping properties that can relate two concepts in a somewhat hierarchical manner. Introducing such information between vocabularies greatly enriches the semantic knowledge already embedded in *legivoc*. Yet, the size of the *legivoc* database, with its (up-to-now) 13 vocabularies sporting each an average number of concepts in excess of 7,000, would make this an expensive undertaking if one were to only use the text-based concept (and alignment) editor provided by *legivoc*.

One possible approach to handle data in such high volumes is to rely on automatic semantic analysis tools and techniques, e.g., machine learning, to suggest possible alignments to the user, an approach akin to the one taken, for instance, in the EU-Cases project4. Given the more homogeneous nature of our data (only vocabularies), we decided to try another route. We designed an extension to legivoc so that any benevolent user can help improve, in an intuitive, friendly and even "fun" way, the budding ontology induced by alignments, thus following approaches such as crowdsourcing (Wikipedia), human computation [1] or "gamification" [2].

[4] `http://eucases.eu`

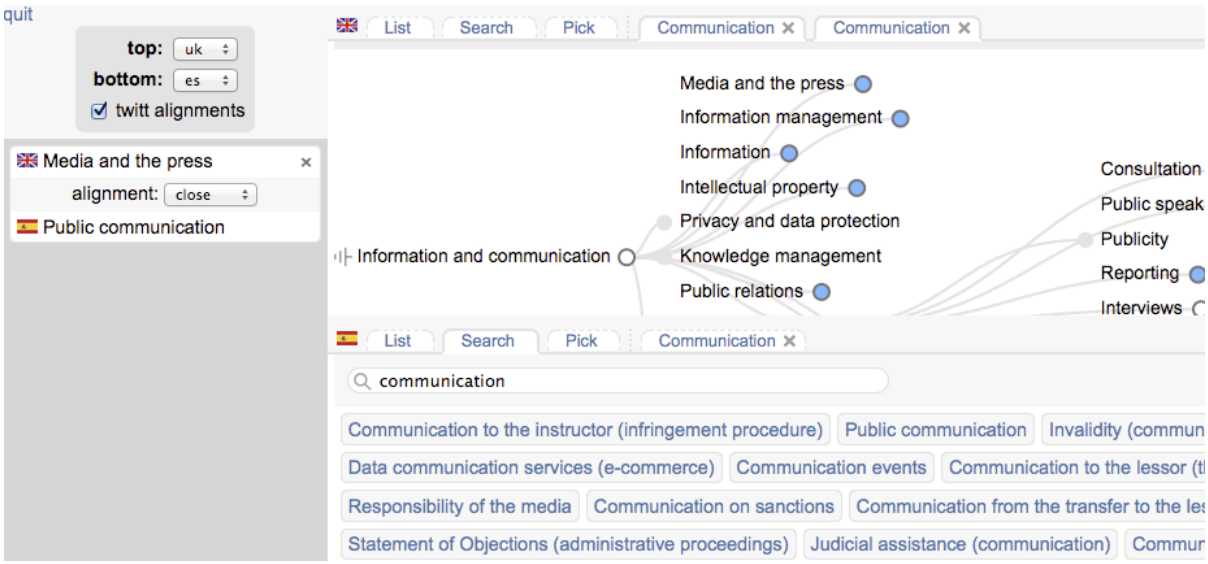


Fig. 5. Aligning the UK and English-translated Spanish vocabularies

With this extension, *legivoc* is now open (in read mode) to any user interested in participating in this ontology-building effort by providing new alignment information; the only user data required is a valid email address and an optional Twitter account. Indeed, all alignments are reported on Twitter, enabling following users to discover in real time when new alignments are being added and by whom, thus creating a sense of community and challenge. New alignments are currently checked a posteriori by experts, to prevent as much as possible the input of possibly incorrect or conflicting information in the *legivoc* database; if need be, malevolent users could be blocked in short order and their actions reverted by a SPARQL command.

There are currently three new modalities for introducing in an intuitive manner new concept alignments involving two different vocabularies, based on how the relevant concepts are found:

- *list*, where concepts are displayed in alphabetic order;
- *search*, where the user inputs keywords to *legivoc*, which returns a list of concepts matching those words;
- *pick*, where the system chooses at random a given concept, and challenges the user to find a related concept in the other vocabulary.

The user then only has to mouse-select one concept in a given vocabulary and drag it on top of the other concept, in the second vocabulary. This process is made even easier and more informed via the use of dendograms, which display in an intuitive manner the hierarchical structure of law concepts (see Figure 5 for a particularly rich example).

A significant illustration of the alignment environment is provided in Figure 5. The top half of the screen is dedicated to the UK vocabulary, in list mode (by clicking

twice on a concept name, here "communication", its partly elided dendogram has appeared), while the second half is in search mode on the Spanish vocabulary. Note that we use the English version of the Spanish law, obtained via MT@EC. The user has already dragged the blue bullet along the word "Media and the press" on top of "Public communication", thus creating an alignment, reported in the left window.

We are currently experimenting on the motivational aspects of our approach. As already mentioned and as graphically indicated by the ticked box dedicated to alignment tweeting, all alignments are tweeted on the `@legivoc` account, in the hope of motivating users. We also intend to use in the future this medium to keep users informed of *legivoc* developments and upgrades. Ultimately, users could also contribute their own tweets on this account, on a limited basis.

4 *legivoc* as a Web Service

We already alluded to the possibility of accessing *legivoc* remotely. This can currently be done in two ways: structured or not.

The first one is via SPARQL commands (see Section 2) sent to a dedicated *legivoc* URL: `legivoc.org/sparql_form`. The first argument, `query`, is a string corresponding to the command to run, while the second one, `db`, is the ISO country code of the vocabulary on which the command is to be run. The output uses the XML format required by the SPARQL specifications. This service allows arbitrary, non-modifying requests to be run on the *legivoc* database, thus providing a powerful and generic API for remote users and servers.

The second access method is via the `legivoc.org/search_concepts` URL. It can be used to perform non-structured, textual searches on the *legivoc* database. The output, which includes the matching concepts URIs, with some additional data, is formatted according to the JavaScript Object Notation (JSON) standard.

Alignments, being encoded within *legivoc* as full-fledged triples, can be retrieved remotely as well as all other data elements.

Obvious security concerns can naturally be raised by such powerful remote access mechanisms. *legivoc* relies on dedicated cookies to ensure proper, fine-grained user rights management; these small pieces of data are transferred and locally stored after performing POST requests at the `legivoc.org/logged_in` URL, with proper name and password parameters. These transferred cookies will have to be passed along subsequent data access requests, ensuring that only registered users are granted access to the system.

In order to validate *legivoc*'s remote access approach, it is currently being tested within the legicoop[5] network for legislative cooperation between the Ministries of Justice of the European Union.

[5] `http://legicoop.eu`

5 Conclusion

We presented a new approach to the alignment input process for the international legal vocabularies stored within the *legivoc* infrastructure. Heavily based on existing W3C standards, it offers both an intuitive and "fun" interactive interface and a remote API. We rely on motivational techniques based on social networks tools such as Twitter to (hopefully) increase the amount of alignment information required to make *legivoc* a success.

Future work will address reasoning over alignment information, e.g., transitive properties or alignment semantics. On the motivational front, it will be mostly driven by the measured effectiveness of the tools used in *legivoc* to fuel the alignment process. In addition to relying on community effects, one could, if need be, envision looking at a higher degree of gamification.

Acknowledgments. We thank Claire Medrala for her help with the implementation of the Twitter interface. We also thank the reviewers for their helping us improve our paper.

References

1. Luis Von Ahn. 2005. *Human Computation*. Ph.D. Dissertation. Carnegie Mellon Univ., Pittsburgh, PA, USA.
2. Sebastian Deterding, Miguel Sicart, Lennart Nacke, Kenton O'Hara, and Dan Dixon. 2011. Gamification using game-design elements in non-gaming contexts. *CHI '11 Extended Abstracts on Human Factors in Computing Systems* (CHI EA '11), ACM.
3. Spyridon Pilos. 2014. European Commission machine translation and public administrations. *Legivoc conference*, Brussels, Belgium. online: `https://www.youtube.com/watch?v=B_rDUisXaB8`
4. Hughes-Jehan Vibert, Pierre Jouvelot, Benoit Pin. 2013. Legivoc - Connecting law in a changing world. *Journal of Open Access to Law*, 1(1)