Developing a semantic legal research interface for the OECD Services Trade Restrictiveness Index

Frédéric Gonzales[1]*, Sébastien Miroudot[1], Thierry Vebr[1]

1: Organisation for Economic Co-operation and Development (OECD), France
**frederic.gonzales@oecd.org*

Abstract

The Services Trade Restrictiveness Index is an OECD policy tool that helps identify regulatory measures that restrict trade. It includes a database of services regulations for 22 sectors in 42 countries, covering legal texts in more than 20 languages. Keeping this database up-to-date on a yearly basis is a priority for the OECD. As the information spans over 18,000 regulations, it is difficult to monitor changes in the underlying data. The aim of a semantic-based legal research assistant is to identify via government web sites, any change in the regulations that are used as a source in the STRI database. In this contribution, we will present first results using natural language processing and semantic enrichment techniques applied to open sources of law in 3 countries (Chile, France and New Zealand) for a set of 50 key regulatory measures affecting trade in services. We will show that this approach has great potential as it could easily be extended to other OECD topics and to other countries for which the adequate data sources are available.

Keywords – OECD Services Trade Restrictiveness Index; Semantic Legal Research Interface; Open Legislation Data; Natural Language Processing; Semantic Enrichment

1 Introduction

The Services Trade Restrictiveness Index (STRI) was successfully launched at the OECD Ministerial meeting in May 2014. It uses a number of indicators to rank countries according to their services trade restrictiveness (Geloso Grosso et al., 2015). Since its last update in December 2015, it includes a database of services trade regulations for 22 sectors in 42 countries and more than 20 languages. This regulatory database covers a high variety of detailed sector-specific regulations in professional, telecommunications, transport, audiovisual and financial services, among others. Keeping this database up-to-date is a priority for OECD member countries. As almost a third of the countries covered by the STRI provide open access to their full national legislative texts (Ubaldi B., 2013), we designed a semantic legal research interface that is able to retrieve legal articles in a faster, more accurate and more exhaustive way, as compared to the original approach where a jurist, consultant or member of the STRI team was using the 'find' option in an html page or a pdf document to identify the same articles (Schweighofer E., 1999). Setting up a standardized approach to analyze different national legal systems in their domestic languages is not without challenges. The most obvious one is how to tackle three legal corpora with a flexible enough research tool. The research assistant presented below offers 2 options: one with predefined search queries, *STRI.Discover*, and another one enabling the user to create its own queries, *Law Search*, which could potentially be applied to topics not related to services trade restrictions.

2 Open Legislation Data

In the pilot phase of the project, we used the Global Open Data Index that collects and presents the current state of open data in 97 countries since 2013 in order to identify the candidates for our research assistant. This independent assessment compares countries over 10 variables, among them: government budget, elections, company register, national statistical office data and national legislation (laws and statutes). In our view, it was then critical to identify countries that would provide open legal data that at least match the following criteria:

- Free to use

- Complete (i.e. an exhaustive set of laws for the country considered)
- Available in bulk download
- Machine-readable format.

The fourth criteria was the most important as it guarantees the possibility for the semantic software Temis Luxid and the match queries to work more efficiently when the corpus consists of millions of legal texts.

From the sample of 13 countries identified as potential candidates and due to time and technical constraints (mainly related to language), three countries only were finally selected for this study: Chile, France and New Zealand. While these three countries fulfill the criteria defined above, they differ in various ways, including the legal system, language, format (all in XML but with different models) and original legal data accessible in bulk download or not.

2.1 Chile

Chile provides open access to its legislation through a SPARQL endpoint. This eases tremendously the exploration of the metadata available in different formats (title, subsection, articles, date of revision in XML among others). The most recently revised regulations were downloaded in XML format using open source software and a web scraping approach, reaching a volume of 280 000 legal texts.

2.2 France

France displays its full national legislation in a bulk downloadable dataset (list of zip files downloadable from an ftp server). It contains the exhaustive database of the French codes, laws, law-decrees, ordinances, decrees in their current, amended and abrogated versions in XML at the article level. The legal corpus used for this study covers the period from 1945 to 2014, including more than 5 000 000 articles (a new update was posted in summer 2016 but was too late to be used in this study).

2.3 New Zealand

New Zealand provides access to its national legislation in XML format at the law level but not in a bulk download. The download of the legal texts was made using web scraping methods with open source software.

3 Methodology

3.1 Data collection

This section describes the general process developed to collect the three national legal corpora and the formatting of the data before the input in the search interface.

Figure 1. Data collection and formatting process.

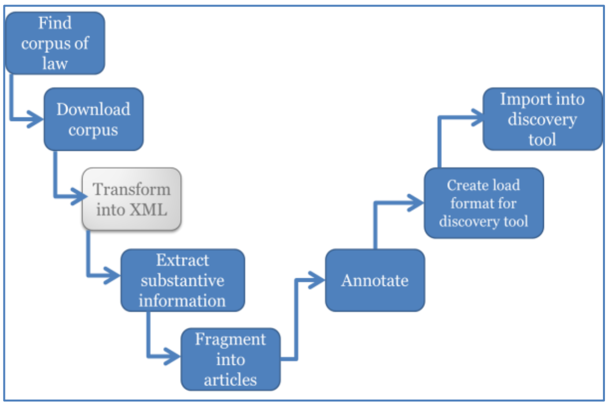

Find corpus of law
Download corpus
Transform into XML
Extract substantive information
Fragment into articles
Annotate
Create load format for discovery tool
Import into discovery tool

Figure 1 presents the detailed steps of our data collection and formatting approach. The first two steps have already been discussed in the country sections above. The “Transform into XML” step is relevant as a general approach covering data not initially available in XML format. In order to guaranty the performance of the semantic treatment, laws are fragmented into articles. The annotation step uses the Luxid annotation server combined with the use of a customized Smart Taxonomy Facilitator cartridge based on the OECD taxonomy enriched by the STRI-specific concepts when necessary. The “Create load format discovery tool” step transforms the initial XML format into standardized XML data for the three countries used by the interface.

2.3 Define legal queries for the research assistant

This section explains how the concepts added to the OECD+STRI taxonomy and the queries are created.

Figure 2. Creation of the queries – Semantic approach

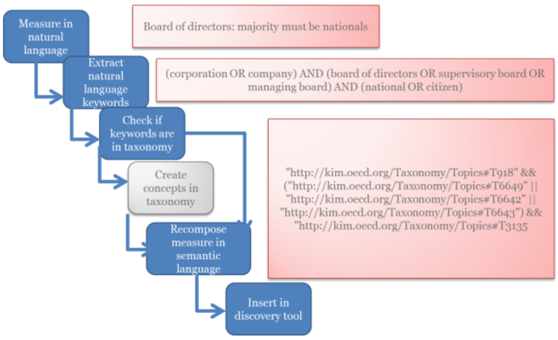

Measure in natural language
Board of directors: majority must be nationals
Extract natural language keywords
(corporation OR company) AND (board of directors OR supervisory board OR managing board) AND (national OR citizen)
Check if keywords are in taxonomy
Create concepts in taxonomy
Recompose measure in semantic language
"http://kim.oecd.org/Taxonomy/Topics#T918" && ("http://kim.oecd.org/Taxonomy/Topics#T6649" || "http://kim.oecd.org/Taxonomy/Topics#T6642" || "http://kim.oecd.org/Taxonomy/Topics#T6643") && "http://kim.oecd.org/Taxonomy/Topics#T3135
Insert in discovery tool

Figure 2 describes the semantic process starting with the STRI barriers in natural language used to create keywords becoming new concepts to be added to the existing OECD taxonomy when they do not exist.

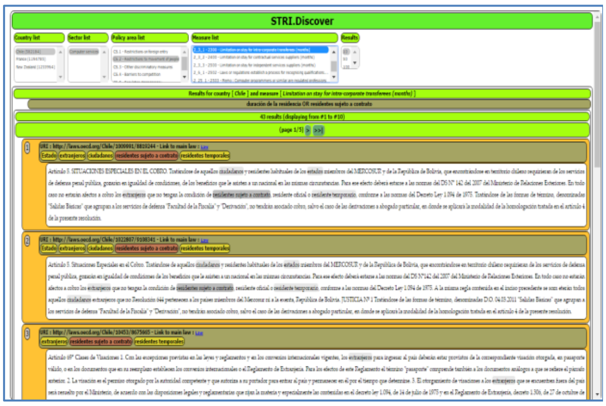
The semantic process and the taxonomy enrichment process are common to both research options. This process uses OECD taxonomy to retrieve information in internal OECD working papers and the official publications database. When missing, STRI related concepts were added to this taxonomy. Another key dimension is the language of the corpus analysed. The OECD taxonomy is available in English and French only. In order to define the Spanish queries for the Chile legal corpus, STRI related concepts in Spanish had to be created. From the perspective of the STRI tool, the addition of a country with an alternative language to English, French and Spanish implies the creation of the equivalent of the STRI taxonomy in this language.

4 A prototype semantic legal research interface for Chile, France and New Zealand

4.1 Option 1 – STRI.Discover

Two web interfaces were created using the same corpus of laws: STRI.Discover, implementing automatic STRI-specific queries, and Law Search, which can be used to search the entire OECD+STRI taxonomy in this corpus.

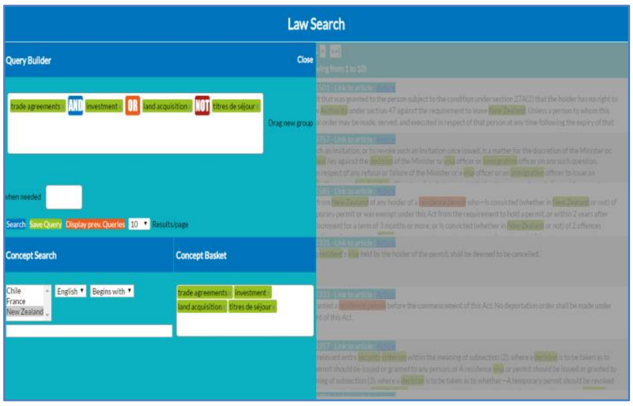
Figure 3. The STRI.Discover web interface


STRI.Discover
Country list
Sector list
Policy area list
Measure list
Results
Results for country [Chile] and measure [Limitation on stay for intra-corporate transferees (months)]
duración de la residencia OR residentes sujeto a contrato
43 results (displaying from #1 to #10)
(page 1/5) > >>|
URI : http://laws.oecd.org/Chile/1009001/8819244 - Link to main law : Law
Estado
extranjeros
ciudadanos
residentes sujeto a contrato
residentes temporales
URI : http://laws.oecd.org/Chile/1022807/9108341 - Link to main law : Law
URI : http://laws.oecd.org/Chile/10453/8675665 - Link to main law : Law

STRI.Discover was the first web interface created. It enables the user to estimate the quality and accuracy of results retrieved by queries defined for more than 50 services trade barriers as listed in the STRI regulatory database. Despite its promising results, the major constraint of this interface is its lack of flexibility as it does not allow the user to create its own queries. Figure 3 displays STRI.Discover results for a search on limitation on stay of intracorporate transferees in Chile. Result #8 listed on page 1 matches the article identified by the STRI team for Chile.

4.1 Option 2 – Law Search

Figure 4. The Law Search web interface.



Law Search allows the creation of personal queries by the user, combining concepts entered in the OECD taxonomy/thesaurus with the three logical operators OR, AND and NOT, as illustrated in Figure 4 above. Hence, the user can define the most accurate queries relative to the topic of interest and not only services trade restrictions.

Finally, for both Chile and New Zealand, the metadata of the articles permit a display of the link to its online web version beside the legal reference of the section of text (e.g. Article #X of Law XX). As a result, the user can directly verify whether the online version has been modified since the implementation of the interface.

5 Conclusion

This study has described a new semantic legal research assistant developed at the OECD and the first results based on legislative data from Chile, France and New Zealand.

The results show how tackling different legal systems (common law and civil code) using three different languages was feasible with one standardized search interface (Law Search). The pilot phase suggests that working with predefined search queries related to services trade restrictions would be useful in future STRI work. In addition, the Law Search interface gives the user the freedom to customize its own queries and test their accuracy when applied to a specific topic such as services trade restrictiveness. It is also a welcomed feature that will help to make the tool more effective.

The ultimate goal of the Law Search interface may now be to deliver a platform for using and enriching big open data, not only legal data as for the STRI project, but also trade agreements, policy papers and research papers etc. We see many potential uses in the context of OECD work.

Although the legal research assistant is still in its pilot phase, it is worth mentioning that no big data tools have been used so far, neither for the data collection and treatment nor for the semantic analysis. Aside from Luxid, all the tools used were open source. Despite what could have been a major constraint for an efficient treatment of millions of text documents, the performance of the interface is more than reasonable. Indeed, this interface works as an OECD internal web tool and successfully provides results in less than 1 second on average. For the production phase, which will involve the expansion to new countries' legislation and other types of corpora such as academic papers and trade agreements, big data architecture will have to be envisaged. We also believe that a more sophisticated interface should be able to monitor legal changes directly online and enable easy management of the links reported in the STRI regulatory database.

More concretely, the next step is to extend the corpus to the following countries: the UK, Spain, Germany, the USA, Finland and Korea.

A further step is to add network analysis (Winkels, R., 2015) between regulations within each domestic legislation and a data visualization layer.

Acknowledgements

We would like to thank our OECD colleagues Bruno Urban and Frédéric Abrazian for their contribution to the data collection and web interface, and Alina Deniau, Eric Espinasse and Barbara Ubaldi for their helpful comments and suggestions.

References

Geloso Grosso, M., Gonzales F., Miroudot S., Kyvik Nordas H., Rouzet D., Ueno A., OECD Trade Policy Paper, No. 177, OECD Publishing

Ubaldi, B. (2013), “Open Government Data: Towards Empirical Analysis of Open Government Data Initiatives”, OECD Working Papers on Public Governance, No. 22, OECD Publishing.

Schweighofer E, The Revolution in Legal Information Retrieval or: The Empire Strikes Back, 1999 (1)The Journal of Information, Law and Technology (JILT).

Hoekstra, R., The MetaLex Document Server - Legal Documents as Versioned Linked Data, pp. 128–143. Springer (2011)

Winkels, R., The Openlaws Project: Big Open Legal Data, Conference paper, February 2015