Towards a Linked Open Data Cloud of Language Resources in the Legal Domain

Author: Patricia Martín Chozas
Advisors: Elena Montiel Ponsoda and Oscar Corcho García

2018
AGRADECIMIENTOS

Una traductora es a la Inteligencia Artificial lo que la velocidad es al tocino...
O eso pensaba yo hasta que conocí a Elena y a Lupe.
A ambas, gracias, de corazón, por descubrirme una inmensidad de conocimiento del que nunca había sido consciente.
A Elena, de nuevo, por toda su paciencia, apoyo y cariño.
A Víctor, por su enorme ayuda. A Oscar, por su amabilidad. A Asun, por recibirme en el grupo.
Al OEG entero, porque todos, sin excepción, me habéis acogido como si llevase aquí toda la vida.
A Elvi, por todos los días.
A mis amigos de siempre, porque siempre es siempre.
A mi familia, por su amor, con todo lo que ello conlleva.
A la velocidad, al tocino, a Pablo.
Gracias, gracias, gracias.
SUMMARY

The use of Semantic Web technologies is progressively increasing since they mean a great help for both machines and humans. In recent years, many institutions and companies are taking the leap to Semantic Web technologies, introducing Artificial Intelligence applications in their processes. In the same way, these organisations are opting for open data formats to publish their datasets, since it translates into a seamless exchange of information with other public or private institutions.

In this scenario, Linked Data emerges as an effort to link related data and suggests best practices for exposing, sharing and actually connecting those pieces of data. Many resources from different domains -geography, life sciences, media, etc.- have been published and connected according to these recommendations, conforming what is known as the Linked Open Data cloud or LOD cloud. Within this cloud, a more specific group of linguistic datasets can be found, identified as the Linguistic Linked Open Data cloud (LLOD cloud).

However, the legal domain is currently underrepresented in the LOD cloud, preventing legal experts from taking advantage of the benefits that interconnecting different types of resources can have.

The work presented here is intended to identify open linguistic datasets of the legal domain -and create them in case they do not exist- and expose them as linked data to contribute to the LLOD cloud. The result of this work will also be part of a Legal Knowledge Graph, that is one of the main objectives of the ongoing project Lynx, a European innovation action to build smart compliance services for multilingual Europe.

With the aim of contributing to the LLOD cloud with legal language resources, the steps proposed by established methodologies for the conversion of linguistic resources to linked data formats (RDF) have been duly followed.

When no resources were available, legal language resources have been generated from scratch, extracting the terminology from legal corpora with automatic term extraction tools.

This work also relied on Semantic Web models (SKOS, specifically) to convert the identified resources, and link them to other available resources in the LOD cloud. The result of this work has been a review of linguistic resources in the legal domain and of the models used to represent those resources in the Web of Data; a set of five terminologies published according to the linked data principles; some recommendations and adaptations in the methodology followed for the RDF conversion process; a preliminary evaluation of term extraction tools and data management tools; and the documentation of all identified and newly-created resources in a public data portal. Accordingly, the outcome of the whole process shapes a first approach to a Linguistic Legal Linked Open Data cloud that will be used to annotate, classify and translate the legal corpora contained in the Legal Knowledge Graph that will be generated in the Lynx project.
RESUMEN

La aplicación de tecnologías de la Web Semántica se está extendiendo progresivamente, ya que representan una gran ayuda tanto para máquinas como para humanos. En los últimos años, muchas empresas e instituciones han dado el salto a la Web Semántica y han introducido la Inteligencia Artificial en su día a día. Del mismo modo, estas organizaciones han comenzado a publicar sus datos en formatos de datos abiertos, pues fomentan el intercambio de información entre otras instituciones.

En este punto nacen los Datos Enlazados, una tecnología que propone pautas adecuadas para exponer, compartir y conectar conjuntos de datos. Muchos recursos de diferentes dominios -geografía, ciencias, multimedia, etc.- ya se han publicado y enlazado mediante estas recomendaciones, y han dado lugar a una nube de Datos Abiertos Enlazados (LOD cloud, en inglés). Dentro de esta nube se encuentra otro conjunto específico de datos lingüísticos: Linguistic Linked Open Data cloud (LLOD). Sin embargo, el dominio jurídico está poco representado en dicha nube, lo que impide que los profesionales de este campo puedan aprovechar las ventajas de interconectar diferentes documentos entre sí.

Este trabajo se centra en identificar conjuntos de datos lingüísticos abiertos del dominio jurídico (y crearlos, si es necesario), para exponerlos como datos enlazados y contribuir así al enriquecimiento de la LLOD. El resultado de este proceso formará parte del Grafo de Conocimiento Jurídico desarrollado en el proyecto Lynx, que pretende ayudar a difundir la información jurídica en la Unión Europea.

Con el objetivo de contribuir a la LLOD con recursos lingüísticos jurídicos, se han seguido metodologías establecidas para la conversión de recursos lingüísticos a formatos de enlazado de datos (RDF). Si no se han encontrado recursos disponibles, se han generado desde cero mediante extracción terminológica de corpus jurídicos con herramientas automáticas de extracción de términos.

Este trabajo también ha utilizado modelos de la Web Semántica (SKOS, en concreto) para convertir los recursos identificados y enlazarlos con otros disponibles en la LLOD. El resultado de este trabajo consiste en una evaluación de los recursos lingüísticos del dominio jurídico disponibles actualmente; un análisis de los modelos utilizados para representar dichos recursos en la Web de Datos; varias recomendaciones y adaptaciones de la metodología para el proceso de conversión de datos; una evaluación preliminar de las herramientas de extracción de términos y de conversión a RDF; la documentación de todos los recursos identificados y generados en un portal de datos público; y un conjunto de cinco recursos lingüísticos publicados según los principios de los datos enlazados y conectados con otros recursos abiertos de la nube de datos enlazados.

De esta forma, el producto de todo el proceso da forma a la primera versión de la nube de Datos Lingüísticos Enlazados del dominio jurídico que se utilizará para anotar, clasificar y traducir los documentos jurídicos que conforman el Grafo de Conocimiento Jurídico que se creará en el proyecto Lynx.
## Contents

1 Introduction .............................................................................. 1  
1.1 Objectives ........................................................................... 4  
1.2 Structure .............................................................................. 6  
2 Foundations .............................................................................. 7  
2.1 Language Resources ........................................................... 7  
2.2 Terminological Resources .................................................... 8  
2.3 Legal Terminology ............................................................. 11  
  2.3.1 Basic Concepts ............................................................. 11  
  2.3.2 Examples of legal terminological resources ................. 12  
3 State of the Art: Language Resources in the Web of Data .......... 17  
3.1 Models to represent Linguistic Data in the Web of Data ....... 18  
  3.1.1 Terminological Resources in RDF ......................... 19  
3.2 Linguistic Linked Open Data Cloud .................................... 22  
3.3 Legal Language Resources ............................................... 24  
  3.3.1 Current Status ............................................................ 24  
  3.3.2 Legal Language Resources in RDF ......................... 25  
3.4 Linguistic Knowledge Graphs ............................................ 27  
3.5 Involved Technologies ....................................................... 28  
  3.5.1 Term extraction technologies ..................................... 28  
  3.5.2 Technologies for RDF modeling and linking ............. 29  
3.6 Current needs ...................................................................... 32  
4 Contribution ............................................................................ 33  
4.1 Identification of existing resources .................................... 35  
4.2 Creation of new resources ............................................... 43  
  4.2.1 Term extraction stage ............................................... 43  
  4.2.2 Term evaluation stage ............................................... 46  
4.3 Conversion into RDF ........................................................ 47  
  4.3.1 URI naming strategy .................................................. 48  
  4.3.2 Modelling ................................................................. 48  
4.4 Linking step ........................................................................ 52  
  4.4.1 Linking Results .......................................................... 53  
4.5 Data portal .......................................................................... 57  
5 Conclusions and future work .................................................. 59  
6 ANNEXES .............................................................................. 63  
6.1 Accepted abstract for “Law via the Internet” Conference. October 2018 ......................................................... 66  
6.2 Accepted abstract for “Encuentros Complutenses en torno a la Traducción” Conference. November 2018 ..................... 71
List of Figures

1  Graphical representation of the Linguistic Legal Linked Open Data cloud........................................... 5
2  Example of term entry extracted from the Black’s Law Dictionary.................................................. 12
3  Example of term entry extracted from the Dudario jurídico de la ONU............................................. 13
4  Example of term entry extracted from the bilingual glossary of the IMF........................................... 13
5  Example of term entry extracted from the monolingual glossary of the IMF...................................... 13
6  Example of term entry extracted from UNTERM................................................................................. 14
7  Example of legal term entry extracted from IATE................................................................................ 14
8  Example of legal term entry extracted from IATE (extended entry).................................................... 15
9  Example of agricultural term entry extracted from IATE................................................................. 15
10 Example of parallel corpora in Spanish, English and German, extracted from EUR-Lex.......................... 15
11 Example of RDF graph represented by SKOS (from SKOS Core Guide). ............................................. 19
12 Graphic representation of the Linguistic Linked Open Data Cloud.................................................... 23
13 Flow diagram showing the stages of the process.................................................................................. 34
14 Datasets represented by domain........................................................................................................ 41
15 Datasets represented by format........................................................................................................... 41
16 Example of candidate terms in Sketch Engine..................................................................................... 45
17 Example of term search in Linguee......................................................................................................... 46
18 Example of the candidate terms........................................................................................................... 47
19 Example of the glossary structure (I).................................................................................................... 50
20 Example of the glossary structure (II)................................................................................................... 50
21 Example of RDF skeleton in OpenRefine (I).......................................................................................... 50
22 Example of RDF skeleton in OpenRefine (II)......................................................................................... 51
23 Example of RDF converted glossary..................................................................................................... 51
24 Suggested sens of the term “erasure”..................................................................................................... 53
25 Term not matched automatically........................................................................................................... 54
26 Too general term not matched............................................................................................................... 54
27 SKOS property not identified................................................................................................................ 55
28 First approach of the Linguistic Legal Linked Open Data cloud.......................................................... 56
29 Example of UNESCO Thesaurus documented in CKAN.................................................................... 57
30 Terminesp entry (included in Terminoteca RDF) modelled with Ontolex................................................ 65
# List of Tables

<table>
<thead>
<tr>
<th>Table Number</th>
<th>Table Title</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>Table of linguistic resources. Adaptation from Montiel-Ponsoda.</td>
<td>8</td>
</tr>
<tr>
<td>2</td>
<td>Comparison of significant terminological resources.</td>
<td>10</td>
</tr>
<tr>
<td>3</td>
<td>Project and vocabularies applied.</td>
<td>21</td>
</tr>
<tr>
<td>4</td>
<td>Archived resources.</td>
<td>39</td>
</tr>
<tr>
<td>5</td>
<td>Set of available language resources identified.</td>
<td>40</td>
</tr>
<tr>
<td>6</td>
<td>Corpora comparison.</td>
<td>46</td>
</tr>
<tr>
<td>7</td>
<td>Examples of term URIs.</td>
<td>48</td>
</tr>
<tr>
<td>8</td>
<td>SKOS properties applied.</td>
<td>49</td>
</tr>
<tr>
<td>9</td>
<td>DublinCore properties.</td>
<td>49</td>
</tr>
<tr>
<td>10</td>
<td>Results of the linking tests.</td>
<td>53</td>
</tr>
<tr>
<td>11</td>
<td>Legend of the LLLOD graph.</td>
<td>56</td>
</tr>
<tr>
<td>12</td>
<td>Correspondence of dataset metadata in JSON.</td>
<td>63</td>
</tr>
<tr>
<td>13</td>
<td>Correspondence of resource metadata in JSON.</td>
<td>63</td>
</tr>
<tr>
<td>14</td>
<td>Term extraction tools comparison.</td>
<td>64</td>
</tr>
</tbody>
</table>
1 Introduction

Law is constantly growing as society evolves. New regulations are created every day, and tasks such as organising, structuring and filtering the documents that describe these regulations comprise great efforts. Consequently, users of law (judges, attorneys, businesspeople and even students) find difficulties when having to cope with the increasing amount of documents involved in these procedures and when identifying the laws that apply.

Five factors causing such difficulties in law management have been spotted [1]:

1. Multiple jurisdictions: in Europe, for instance, laws are enacted at different levels (municipal, regional, national and European). The process of identifying legal requirements in such cases involves an enormous amount of time.

2. Volume: because of the reasons stated above, it is almost impossible to be aware and updated on the legislation. The European legislation, for instance, is 170,000 pages long.

3. Accessibility: on-line law portals ease the access to legislation, but the current situation still has many issues to overcome. Only a few official institutions support such portals, since creating them is expensive and not every government can afford it. Those that can afford it encounter many problems when homogenising documentation (in terms of format, content, language, jurisdiction...)

4. Updates and consolidation: in many occasions, there is not a proper version control of the laws. Some laws state which articles are modified, but others do not. In such cases, the decision of applying a certain version of a law only depends on the judge in charge, and this can lead to inequalities.

5. Vague classification of laws: depending on the product to which the laws are applied, these can belong to one domain or another. Some problems arise because sometimes laws are applied to the wrong domain.

Artificial Intelligence (AI) can be a great help in solving these problems, and more specifically Natural Language Processing (NLP) and Semantic Web.

By combining both fields, technologies such as Linked Data are born, creating in turn applications that provide an easier and faster access to documents and information through the generation of links.

Currently, many resources in the legal domain can already claim to be connected, like any HTML or XML documents with hyper-references to other documents. However, a very efficient way to describe connected resources on the Web relies on W3C specifications of the Semantic Web, such as RDF [2], RDFs, OWL [3] or SKOS [4]. Linked Data is a particularly sound manner of publishing RDF resources that need to be structured according to the Linked Data Principles [5], namely: entities should be identified via unique URIs; URIs should be HTTP URIs, follow standard web protocols, return useful information about the resource, and contain links to other
related resources. This way of exposing resources allows to move from connections or links at the document level to connections between the data items mentioned in a document, which results in a much more specific way of establishing links between data.

Some datasets that are published as Linked Data are part of the *Linked Open Data (LOD)* cloud, a diagram representing connected linked data resources. These two concepts should not be used in-distinctively: Linked Open Data is Linked Data published with an open license that allows its free access and use [6]. The LOD diagram exposes the different linked resources sorted by domain. Currently, the LOD cloud categorises the data resources contained there into nine domains: Cross-Domain, Geography, Government, Life Sciences, Linguistics, Media, Publications, Social Networking and User Generated resources.

Apart from this classification per domain, another sub-cloud has been identified that gathers resources of the same nature. It is the *Linguistic Linked Open Data* cloud or *LLOD*, a cloud of linguistic resources that follow the same criteria of openness, availability and interlinking as the LOD, as stated in its website\(^1\).

Since this thesis is focused on legal language, only the Linguistic Linked Open Data cloud, which is restricted to datasets in the linguistic domain, will be considered. However, language resources in the Linguistic LOD do not only belong to the legal domain. In fact, there are only a few that can be considered as legal language resources. Those datasets will be analysed to check if they can be reused in this project.

On the other hand, apart from evaluating the legal language resources, this work has also studied the legal language itself, since it has its own peculiarities [7]:

- Sentences are usually very long and complex.
- Punctuation marks are scarce, which slows down the reading and comprehension.
- In many occasions, expressions in foreign languages (normally, in Latin) are used; previous knowledge of these languages is required.
- Even in the same language, it is very common to see rare and complex expressions and words only used in legal language. They can be sorted out into four categories:
  - Legal terms of art: technical words phrased with exact meaning that cannot be replaced by other terms (they usually appear defined in a legal dictionary or glossary).
  - Legal jargon: terms or expressions used by lawyers, often archaic and obsolete general words, that usually have a correspondence in plain language to ease non-law users understanding.

\(^1\) [http://linguistic-lod.org/llod-cloud](http://linguistic-lod.org/llod-cloud)
– Terms from the general language with a different meaning in the legal domain.
– Terms from the general language that keep their ordinary meaning but are applied in rare contexts in the legal domain.

In addition, and taking into account that we live in a globalised world in which law decisions taken in one country end up influencing the decisions taken in another country, the issues mentioned above would multiply by the number of languages involved. **Multilingualism** is, therefore, another difficulty in law management, and those peculiarities of legal language are added to the existing general difficulties of each language. For instance, declinations of German, unclear grammar of English, etc.

In particular for legal Spanish, nine issues have been identified that generate ambiguous interpretations of law [8]:

- The use of comma before conjunctions and some prepositions
- Long distance between subject and verb, specially when representing collective entities
- Excessive use of passive structures
- Excessive use of anaphoric pronouns
- Excessive use of relative pronouns
- Excessive use of gerund
- Excessive use of adverbs
- Use of the structure *and/or*
- Concatenation of subordinate sentences

Because of all the factors mentioned above, not only managing but also understanding law is a task usually delegated to lawyers and law firms, and not undertaken by individuals without the required knowledge or companies without their own team of lawyers. They need to retrieve information from many disparate sources, which are often published in various formats by different institutions at national red and international levels. This problem is heightened when we consider a context as the European one in which in a “Single Market” several jurisdictions **coexist**. This means that when companies want to go outside their local markets they face a foreign jurisdiction, unfamiliar practices, and all of it expressed in a language other than red that of their own.

Multinational corporations and worldwide enterprises are used to this kind of processes and they count on the help of their own lawyers, who are usually up to date with the latest changes in regulations. However, for **small and medium size**
enterprises (SMEs) that are trying to expand and sell their products abroad the management of such tasks can turn into an insurmountable obstacle.

Aimed at helping SMEs overcome all the hindrances that may arise when trying to comply with international law in the context of the European Union, and also lawyers in the legal department of enterprises, the Lynx project has been launched. Lynx² is an Innovation Action funded by the European Union’s Horizon 2020 research and innovation programme under grant agreement no 780602. It started on the 1st of December 2017 and has a duration of 36 months.

The goal of Lynx project is to create a Knowledge Graph of legal and regulatory data to facilitate the access to information from different jurisdictions, languages and domains red through a set of so-called compliance services. A Knowledge Graph is understood as a structure that represents information, where entities are represented as nodes, their attributes as node labels and the relationship between entities as edges [9].

In particular, this graph will contain data covering the domains of the three business cases represented in the project: labour law, data protection and industrial norms and standards. Furthermore, not only this will be useful for SMEs and companies in general, but every European citizen will be able to benefit from this collection of interlinked legal data, that will be open.

The contribution of this work, that will be described in the next section, will be part of the creation of the Legal Knowledge Graph, building the linguistic foundation that will support it.

1.1 Objectives

The general purpose of this work is the creation of a Linguistic Legal Linked Open Data cloud as part of the Legal Knowledge Graph that is to be built in Lynx project. This graph will interlink public and private legal resources, metadata, standards, general open data and language resources. The Linguistic Legal Linked Open Data cloud is intended to contribute to the language resources to be contained in the Legal Knowledge Graph. With this aim, a series of individual objectives have been identified and pursued:

- The first objective of the work presented here is the identification of existing language resources of the legal domain and its classification according to accessibility and reusability.

- The second objective is the analysis of the format in which such resources are published to evaluate which are the most useful datasets for the project and choose the best manner to convert them into linked data formats.

- The third objective is to create a new set of legal resources, specific for the project, from legal corpora provided by the partners of the Lynx consortium.

² http://lynx-project.eu/
• The **fourth objective** is to convert both sets of resources into RDF: the already existing assets that have been identified (first objective) and the new language resources extracted from Lynx corpora (third objective).

• The **final purpose** is contributing to the Linguistic Linked Open Data cloud (LLOD) with linguistic resources of the legal domain, that will be eventually gathered in the Linguistic Legal Linked Open Data cloud (LLLOD) (see Figure 1).

---

**Fig. 1:** Graphical representation of the Linguistic Legal Linked Open Data cloud.
1.2 Structure

The work presented in this document is structured as follows:

- Section 2 contains a theoretical framework intended to put this work in context and to support it with a linguistic foundation. This section gathers information about general language resources (terminologies, specifically), legal language resources and examples of each type.

- Section 3 is dedicated to the State of the Art, which is divided in several subsections: language resources in the Web of Data; models to represent Linguistic Data and resources already available; Linguistic Linked Open Data cloud, legal language resources and those available in RDF; linguistic knowledge graphs; technologies applied in this work and the current needs derived from the State of the Art.

- Section 4 collects the whole contribution and processes that took part in this project, namely, identification of existing resources, creation of new resources, conversion into RDF and linking process. It also contains a section devoted to the data portal where all the resources handled in this project have been documented.

- Section 5 includes the conclusions drawn from this work. It also contains the next steps that are to be taken in Lynx project and some issues that should be solved to improve the current situation of the Semantic Web technologies.

- Finally, Section 6 shows additional tables to clarify the information given along the document, as well as two conference paper proposals: one of them accepted and the second one still under review.
2 Foundations

2.1 Language Resources

Connecting legal documents from across the European Union is an ambitious task that requires a strong foundation on language resources to support a structured organisation of all these documents and the data they contain. For the purposes of this thesis the terms language resources and linguistic resources will be used without distinction.

There is no consensus about the definition of a language resource, neither about the kind of assets that are covered by this term. However, the following are considered common methods of organising linguistic information [10]:

- Glossary: a collection (lexical list, catalogue) of specialised terms and their meanings (Example: Glossary of Abbreviations developed by UNESCO\(^3\)).
- Lexical Database: a resource that contains pieces of linguistic knowledge stored in a systematic way as data elements so that a computer application can access them (Example: DANTE, a lexical database for English\(^4\)).
- Dictionary: a record that organises lexemes of a language in form of lemmas following an alphabetical order and where their meaning is explained (Example: DLE, dictionary of Spanish\(^5\)).
- Thesaurus: a controlled list of semantically and generically related terms that cover a specific domain of knowledge hierarchically structured (Example: TheSoz, German thesaurus for social sciences\(^6\)).
- Lexicon: also known as lexical knowledge bases, they are considered here as networks with information about words and their contexts. They usually appear in groups of synonyms that provide short definitions and relations between them. Lexicons seem very convenient to disambiguate; they are useful to assign the right term to a given context (Example: WordNet\(^7\)).

Bearing in mind the structure of these resources, specificity, representation of the information and end users, together with the purpose of this thesis, it seems necessary to add one more type of language resource to the list: terminologies. Terminologies and glossaries are often considered as equivalent, since both are catalogues of specialised terms (they can also contain definitions and translations). However, while a glossary may define specialised terms from several domains, a terminology will always collect terms sorted by domain (for instance, legal terminology). Also, glossaries use to be composed just by the term and the definition. Terminologies

\(^3\) http://www.unesco.org/education/educprog/brochure/032.html
\(^4\) http://www.webdante.com/
\(^5\) http://dle.rae.es)
\(^6\) http://lod.gesis.org/thesoz/de.html
\(^7\) http://wordnetweb.princeton.edu/perl/webwn
contain this kind of information as well, but also term variants, synonyms, usage
notes, context and sources.

Table 1 shows a comparison of the linguistic resources already mentioned. Such
table has been modified to include terminologies in the study.

<table>
<thead>
<tr>
<th>Criteria</th>
<th>Glossary</th>
<th>Database</th>
<th>Dictionary</th>
<th>Thesaurus</th>
<th>Lexicon</th>
<th>Terminology</th>
</tr>
</thead>
<tbody>
<tr>
<td>Organization</td>
<td>Alphabetical order</td>
<td>Alphabetical order</td>
<td>Alphabetical order</td>
<td>Semiantically</td>
<td>Related entries</td>
<td>Alphabetical order</td>
</tr>
<tr>
<td>Semantic information</td>
<td>Definition</td>
<td>Definition</td>
<td>Definition</td>
<td>Hierarchical</td>
<td>Related entries</td>
<td>Definitions</td>
</tr>
<tr>
<td>+ Other info</td>
<td>+ Pos</td>
<td>+ Etymologies</td>
<td>+ Derivation</td>
<td>+ Translations and terminological relations</td>
<td>+ Usage</td>
<td>+ Translations and terminological relations</td>
</tr>
<tr>
<td>Physical format</td>
<td>Paper</td>
<td>Electronic format</td>
<td>Paper</td>
<td>Electronic format</td>
<td>Paper</td>
<td>Electronic format</td>
</tr>
<tr>
<td>+ Electronic format</td>
<td>+ Paper</td>
<td>+ Electronic format</td>
<td>+ Electronic format</td>
<td>+ Paper</td>
<td>+ Electronic format</td>
<td></td>
</tr>
<tr>
<td>Domain of knowledge</td>
<td>General</td>
<td>General</td>
<td>General</td>
<td>Specific</td>
<td>Specific</td>
<td>Specific</td>
</tr>
<tr>
<td>+ Specific</td>
<td>+ Specific</td>
<td>+ Specific</td>
<td>+ Specific</td>
<td>+ Specific</td>
<td>+ Specific</td>
<td>+ Specific</td>
</tr>
</tbody>
</table>

Tab. 1: Table of linguistic resources. Adaptation from Montiel-Ponsoda.

The latter becomes one of the most important within this context since termi-
nologies are built to represent information of a certain domain or research area, in
this case, the legal domain.

2.2 Terminological Resources

The role of terminology is specially important in this context, since legal terms must
be accurately defined so that they can be used in the description and annotation
of legal documents, and it allows the establishment of correspondences or links
amongst the annotated documents. Terminology is mainly understood as the field
of study specialised in identifying, analysing, describing, and relating terms. It is
an interdisciplinary field that deals with concepts and their representations.

This section will be divided in three parts that contain information about the
main features regarding terminology:

- Terminologies are domain specific. This feature is what makes them such a
  valuable asset.

- Modern terminology theories claim that the starting point in terminology
  study are the terms in a text, and that by analysing them in context we
  can derive their meaning or the concept behind. We will refer to this in more
  detail below.

- Terminology work comprises many different activities applied to several fields
  of study, and different terminological resources have been created accordingly.

Domain Specificity

One of the main features of terminology is that it will always be domain specific. For
this reason, a term is defined as a word or a combination of words that have
specific meaning in a certain domain. This means that a term may represent
two different concepts in two different domains. For instance, the word “court” does not have the same meaning when used in a legal context than when talking about basketball. Thus, terminology is essential to avoid confusing information given by polysemic words, since a terminology will usually include the concept and definition relevant to one specific domain. That is why a term should not be used to represent more than a concept in within the same terminology (Foo, 2012).

**Terminology Theories**

The purpose of this section is to give an overview of how terminology originally understood and how this understanding has evolved over time with the aim of strengthening the theoretical framework of this thesis.

The *General Terminology Theory (GTT)*, promoted by Eugen Wüster (1991), tries to make specialised knowledge universal through the standardisation and normalisation of terms. This theory is based on the idea of a unique concept that can be represented by several terms, depending on the context. Concepts pre-exist terms; this means that the concept exists before term, so that the understanding of a concept can be considered as language independent. Therefore, this theory is founded on objectivism, on the assumption that the perception of a concept is not related with the human observation and experience.

The latter premise is precisely the reason why modern alternative theories were originated. They all have in common a shared understanding of terms as being the ones that provide access to specialised knowledge (to the concept) and being studied and described according to their behaviour in the context in which they appear.

- The *Communicative Terminology Theory (CTT)* puts the focus on the fact that variation is also natural to specialised communication and not only typical of general language, and it states that terms are not isolated, but they have to be analysed according to the relations between and amongst them [11]. In the same way, the definition of the terminological units does not only depend on the conceptual meaning, but also on pragmatical facts such as the context of use. This postulate is, therefore, opposite to the GTT.

- The *Sociocognitive Terminology Theory (STT)* declares that the traditional approach that is based on the definition of concept is too restrictive [12]. In contrast, the STT states that the world is understood by cognitive models, and it proposes the adoption of units of understanding instead of concepts. Therefore, terminological description is based on the use of terms in a given context, but never on isolated and independent entities.

- The *Frame-Based Terminology Theory (FTT)* shares many features with the previously mentioned theories [13]. One of the main ideas of this theory is that thanks to our background understanding of a domain, a cognitive frame is created and a given term can be better understood in this framework. Another fundamental assumption of this theory is that the definition of concepts in a given domain depends on the task that is going to be developed: this
means that the use of one specific term is object oriented. Also, it proposes to reduce the differences between terms and words, considering the study of *textual knowledge units*.

For the purposes of this thesis, **modern terminology theories** have been considered, in the sense that a **less restrictive approach** seems more appropriate for the identification and definition of legal terms, since **law and its applications are constantly evolving**.

### Examples of terminological resources

The importance of terminology has increased over time due to its multiple advantages and applications. Many projects focused on improving terminological resources have arisen at national and European level.

The *Terminology Coordination Unit of the European Parliament (TermCoord)* is in charge of extending the correct management and good practices of terminology work, and it offers access to the terminology of the European Union through EurTerm, the interinstitutional terminology portal⁸.

TermCoord also coordinates **IATE (InterActive Terminology for Europe)**⁹, the multilingual terminological database of the European Union. IATE is available through an open access platform and it constitutes the most important terminological reference for translators and language users in Europe.

On top of that, TermCoord has contributed to the enhancement of Semantic Web technologies by supporting the **upgrade of IATE to a knowledge base**: while a *terminological database* contains isolated terms without semantic interconnectivity, in a *knowledge base* these terms are structured in a logical way, building a 3D network that presents concepts that have linguistic and cognitive information associated [14].

Some of the key terminology initiatives in Spain are the *TERMCAT* catalog, supported by the Centre of Catalonian Terminology, and *Terminesp*, supported by the Spanish Association of Terminology (AETER).

To give an overview about these instances of terminological resources, a comparative analysis of the three resources mentioned has been performed.

<table>
<thead>
<tr>
<th>Features</th>
<th>TERMCAT</th>
<th>Terminesp</th>
<th>IATE</th>
</tr>
</thead>
<tbody>
<tr>
<td>Level</td>
<td>Regional/National</td>
<td>National</td>
<td>European</td>
</tr>
<tr>
<td>Language</td>
<td>ca, en, es, de, fr, it</td>
<td>de, en, es, fr, it, la, sv</td>
<td>bg, cs, da, de, el, en, es, et, fi, fr, ga, hr, hu, it, la, lt, lv, mt, nl, pl, pt, ro, sk, sl, sv</td>
</tr>
<tr>
<td>Access</td>
<td>Open</td>
<td>Open</td>
<td>Open</td>
</tr>
<tr>
<td>Domain</td>
<td>Multidomain</td>
<td>UNE Spanish Standards</td>
<td>Multidomain</td>
</tr>
<tr>
<td>Format</td>
<td>Online Portal/RDF</td>
<td>Online Portal/RDF</td>
<td>Online Portal/RDF</td>
</tr>
</tbody>
</table>

Tab. 2: Comparison of significant terminological resources.

---

Table 2 shows differences and similarities amongst these resources. Some relations are obvious: language availability differs depending on the coverage. However, the most interesting feature of the previous table (in this context) is the format in which the tools resources are published. All of them are available through an online portal, but they are also structured in RDF and published as Linked Data (more information about such resources will be given in Section 3.1.1). Moreover, TERMCAT and Terminesp are part of the TerminotecaRDF project, an initiative to convert terminologies available in Spain into RDF and publish them as linked data.

2.3 Legal Terminology

2.3.1 Basic Concepts

General concepts are thought to make people clearly understand what their represent. On the contrary, legal terms are usually defined by a legislator, and they have an authoritative fixed meaning. Based on this, four techniques to define a legal term can be inferred: (1) not defining it (leaving its lexical definition), (2) defining a term by its context, (3) defining a term by reference, (4) defining a term providing an explicit definition [15]. This means that every terminology will use one or several of these four definition techniques.

However, when it comes to multilingual terminologies, some important remarks should be pointed out. It cannot be assumed that two legal terms from different legal systems that directly translate are equivalent. Each term may adopt a different meaning in the legal framework of each country, which is permeated by the history, the culture and the traditions of that country. Sometimes, to establish a link between two legal terms, a comparative approach is used with the objective of building a bridge between two legal systems, but in no case, to create a one-to-one equivalence [16].

One of the main problems with legal terminology appears in the translation process. Legal translation involves more culture-specific components than other translation fields such as technical or medical translation. This is caused by the system-bound nature of legal terminology: legal terms are based on a national legal system. Legal systems follow their own principles and patterns, and they have been created to comply with the needs of a concrete country, which means that there will always be incongruity between legal terms of different national systems.

To translate a legal term, an equivalent that meets the properties of the original term must be found. It needs to reference the right legal system and be accepted by the community. Such situation may bring several complications to legal translators. Luckily, there are some lexical resources available that can ease this task [17]:

- Monolingual legal dictionaries: they provide definitions of legal concepts that constitute conceptual networks of a single legal system in one language. With this information in source and target language, the translator should be able to match the source term with the corresponding equivalent in the target
language. To accomplish this step, bilingual dictionaries also provide a great help.

- Bilingual legal dictionaries: they provide equivalents for legal concepts in a target language. Entries have a limited degree of detail but combined with monolingual dictionaries they help erase inconsistencies in translation.

- Comparable and parallel corpora: case law, court decisions and any other official and trustable legal documents are particularly useful for discovering the linguistic context in which equivalent terms and expressions are used in original and target language.

Summing up, in this section we have provided a review of different types of legal linguistic resources, both terminological and lexical, and have seen how they can be used in complementary manner.

### 2.3.2 Examples of legal terminological resources

This section gathers examples of the most common legal language resources used in the legal translation field. The list has been created by conducting a survey of legal translators to get information about the resources used in their daily translation activities.

*Black’s Law Dictionary* is a monolingual legal language resource, written in US English, that has become the most important reference for legal translators [17]. One of the key features of this resource is that each entry provides links to other related terms, alongside with information about other senses, usage contexts, authorities applying the term, etc. The following image shows an example of a terminological entry exposed in this resource.

![Fig. 2: Example of term entry extracted from the Black’s Law Dictionary.](image)

With regard to bilingual dictionaries, several English-Spanish glossaries containing information on different legal subdomains have been widely used amongst language professionals and they are now very valuable assets for Spanish translators.
One of these resources is the *Dudario jurídico de la ONU*¹⁰, a dictionary developed by the translation department of the United Nations in New York, and it gives information about the correct usage of a term and the different connotations that it may have, depending on the context. The following figure shows an example of term entry, its possible translations depending on the context and usage examples.

**Clause**

Una cláusula es una disposición de una ley, tratado, acuerdo o contrato. En español, usamos en general “disposición” o “artículo” (si lo fuera) cuando se refiere a una ley, y “disposición”, “artículo” (si lo fuera) o “cláusula” para los tratados, acuerdos o contratos [ejemplos: cláusula de la nación más favorecida, cláusula de resolución, cláusula de arbitraje].

Fig. 3: Example of term entry extracted from the *Dudario jurídico de la ONU*.

Other relevant resources are those developed by the *International Monetary Fund (IMF)*: one bilingual glossary (English into Spanish) that contains terms and a set of possible translations¹¹ (Figure 4), and another smaller monolingual glossary (Spanish) that contains definitions and usage information¹² (Figure 5).

<table>
<thead>
<tr>
<th>English</th>
<th>Spanish</th>
</tr>
</thead>
<tbody>
<tr>
<td>tax audit</td>
<td>inspección de hacienda</td>
</tr>
<tr>
<td></td>
<td>auditoría impositiva</td>
</tr>
<tr>
<td></td>
<td>auditoría fiscal</td>
</tr>
<tr>
<td></td>
<td>inspección tributaria</td>
</tr>
</tbody>
</table>

Fig. 4: Example of term entry extracted from the bilingual glossary of the IMF.

**Convenio Constitutivo (Articles of Agreement).** Tratado internacional en el que se establecen los fines, principios y estructura financiera del FMI. El Convenio, que entró en vigor en diciembre de 1945, fue redactado por representantes de 45 naciones en una conferencia celebrada en Bretton Woods, New Hampshire.

Fig. 5: Example of term entry extracted from the monolingual glossary of the IMF.

As per terminological databases, the *United Nation Terminology Database (UNTERM)*¹³ provides terminology and nomenclatures used in the work of the United Nations.

---

¹⁰ https://onutraduccion.wordpress.com/pref/dudario-juridico/
¹³ https://unterm.un.org/UNTERM/portal/welcome
Nations in the six UN official languages, and also some entries are available in German and Portuguese. This portal extracts terminological information from several datasets, allowing the user to choose their preferred source. Moreover, each entry is linked with the original document in which it has been applied, as it appears in Figure 6.

On the other hand, IATE\textsuperscript{14}, that was mentioned in the previous section, also provides terminological information on European legislation in more than 20 European languages. IATE also contains terms from other domains such as sciences or agriculture. As an example, Figures 4 and 8 show a terminological entry from English into Spanish in the legal domain and Figure 9 shows a related term applied to the agricultural domain.

\textsuperscript{14}http://iate.europa.eu
2.3 Legal Terminology

Lastly, the following asset has also been developed by the European Union and, although it is not exactly a terminological database, it has become a fundamental source of reference for legal language users. **EUR-Lex** is a corpus that contains EU public documents (regulations, directives, decisions, etc.). The tool that allows to browse it, offers the user a bilingual view of parallel corpora, allowing translators and terminologists to check for a specific term in the right context (see Figure 10).

**Fig. 8:** Example of legal term entry extracted from IATE (extended entry).

**Fig. 9:** Example of agricultural term entry extracted from IATE.

**Fig. 10:** Example of parallel corpora in Spanish, English and German, extracted from EUR-Lex.
3 State of the Art: Language Resources in the Web of Data

The World Wide Web Consortium (W3C) works to improve and support the growth of the Web. Much of the content of the Web is intended for human consumption; however, if the data are not published in an organised and machine-readable format, machines and computers will not be able to process them. Therefore, Artificial Intelligence applications intended to help the user access the information easily will not be able to process this information. For this reason, part of the efforts of the W3C to organise the content of the Web lies on Semantic Web technologies, which make reference to any technique used to structure the content of websites for machines and computers to manipulate it, helping simultaneously with the enrichment of the Web of Data.

The Web of Data goes beyond the Web of Documents: in the Web of Data, not only the documents are connected, but the information contained in these documents is also linked with related content. Semantic Web technologies rely on hypertext links, a powerful tool to share information, since “anything can link to anything” [18].

However, not every type of document and format is relevant for the purposes just described. Information shall be structured in a machine readable format that permits exposing resources that contain data. It is at this point when the RDF format comes in, becoming the most common method to publish data in the Semantic Web.

Resource Description Framework (RDF) is a model that supports the concept description, the representation of information and the interchange of data on the web. RDF is based on triplets, subject-predicate-object expressions that are used to make declarations about the resources. Its simple syntax makes it possible to expose and share structured and semi-structured data. As already stated in the introduction, this thesis is aimed at reviewing legal language resources in RDF so they can be eventually published as Linked Data [19].

Any content considered as Linked Data needs to comply with the following principles of Linked Data [6]:

- Entities should be named via unique URIs.
- These URIs should be HTTP URIs and follow standard web protocols.
- These URIs should return useful information about the resource.
- They should contain links to other URIs pointing at related resources.

An example of this principles is the following:

http://tbx2rdf.lider-project.eu/data/iate/competence+of+the+Member+States-en

In this example, the URI acts as a unique identifier of the lexical entry Competence of the Member States, contained in IATE. This URI returns information about
the term, it is also linked with related URIs and it will never be used to point at two different pieces of information.

3.1 Models to represent Linguistic Data in the Web of Data

A large amount of language resources can already be found across the Semantic Web. Such datasets are represented with various models, depending on their structure, objective and content, amongst other features. The most relevant for this thesis are listed as follows:

- **LIR model**: Linguistic Information Repository is proposed to solve issues related with multilingualism in ontologies, considering the localisation of the ontology terminological layer, without modifying the ontology conceptualisation. LIR provides linguistic information necessary for the localisation of the ontologies and consequently offering unified access to multilingual data [20].

- **Lexinfo**: this model is implemented as an OWL ontology, and it is intended to associate linguistic information to elements in an ontology. It works as a declarative model to represent and share ontology lexica [21].

- **lemon**: Lexicon Model for Ontologies, usually known as lemon, was specially created to represent lexical information from ontologies in the Semantic Web. This is composed by a monolingual core lexicon that contains lexical entries (each one representing a specific term) that can be either word, phrases or part of words. This model also represents additional information such as term variants (abbreviations...), lexical senses, lexical forms, etc. Therefore, the core elements of the lemon model are: lexicon, lexical entry, form, representation, lexical sense, reference, property, frame and argument and component [22].

- **Ontolex**: Ontolex is the result of the lemon evolution supported by the W3C Ontology-Lexica Community Group\(^{15}\), and it is intended to represent the lexicon-ontology interface. The mappings between the lexical entry and the ontology are represented by the property ontolex:reference. A great improvement of this model is the Ontolex vartrans module that allows to represent term variants and translations. It also provides representation for relations amongst senses and amongst forms [23] [24].

On the other hand, **SKOS: Simple Knowledge Organization System** is aimed at representing the structure of organization systems such as thesauri and taxonomies, since they share many similarities. It is widely used within the Semantic Web context since it provides an intuitive language and can be combined with formal representation languages such as the Web Ontology Language (OWL). SKOS XL works as an extension of SKOS to represent lexical information [25].

\(^{15}\)https://www.w3.org/community/ontolex/
SKOS considers taxonomies, thesauri and any other controlled vocabulary as “concept schemes”, including glossaries and terminologies, and they can be described with the class skos:conceptScheme.

Each concept scheme has top concepts, that are used for datasets that present broader and narrower relations. skos:hasTopConcept represents a top-level concept that is more general than the rest of the concept in the scheme. For instance, an accurate top concept for the example below (Figure 11) could be “Economy”.

Since SKOS is widely used to represent hierarchical relations, concepts can also have “broader terms” and “narrower terms”, just like a person can have parents and children. These relations are represented with properties skos:broader and skos:narrower.

Regarding the naming strategy, concepts in SKOS have a unique identifier to which the skos:concept property is applied. To show the written representation of the concept, the properties skos:prefLabel and skos:altLabel are used to show the preferred and the alternative denominations respectively (see Figure 11).

SKOS also enables the representation of term definitions, with the property skos:definition, and general term notes (usage advices and additional information), with the property skos:note. These properties are not shown in the example, but they are relevant since they have been applied in this work.

![Fig. 11: Example of RDF graph represented by SKOS (from SKOS Core Guide).](image)

### 3.1.1 Terminological Resources in RDF

Probably, one of the most relevant works in the state of the art of this work is the conversion of the TBX version of IATE into RDF, following the lemon model [23]. Apart from IATE, the work also dealt with the European Migration Network glossary, and the consequent linking of both resources.

The terminologies were organised in a graph, and the terms were represented with broader term and narrower term relations.
The process described in that paper to transform terminologies into RDF seems quite relevant for the purpose of this thesis, since both projects share several similarities: as stated before, the source dataset mentioned in the paper were originally in TBX format, while for this thesis TBX, CSV and XML files have been handled.

This is why it is worth mentioning that they have applied the SKOS vocabulary to represent term entries, by using the property `skos:concept`. Since terminologies can also be regarded as knowledge organization systems, it seems quite pertinent the use of SKOS concepts to represent terminological concepts. For this reason, `skos:concept` property has been used to represent terminological entries in this project.

They have used the Ontolex model to represent lexical information about the entries. This model allows the creation of one lexicon per language and represent translation equivalents by linking the entries in different languages to the same `skos:concept` through the `ontolex:reference` property.

After converting these resources into RDF, the next step was, of course, the linking process, where the entries of both IATE and EMN were linked by the `skos:concept` property to the corresponding lexical entries in different languages.

As a result of the whole process, each terminological concept in IATE was transformed into a `skos:concept`, and for each available term in EMN a `LexicalEntry` was created.

The resulting automatic terminology conversion tool, named as TBX2RDF was presented as a website application\(^\text{16}\). It is open and offers the following functionalities:

- Translate from a TBX to an RDF file.
- The application accepts a TBX file as an input and returns an RDF document (and a description of the errors encountered if any).
- ReverseTranslate would accept RDF as input and return TBX as output (not implemented).

Another of the most important projects in Spain is Terminoteca RDF in which several multilingual Spanish Terminologies were presented as a collection of datasets published as Linked Data [26] [27].

This work is particularly pertinent since it gathers two datasets, and one of them is also involved in this thesis: Terminesp, a multilingual terminological database developed by the Spanish Association for Terminology; and terminological glossaries from the Terminología Oberta service of the Catalan Terminological Centre (TERMCAT).

The latter can be open accessed, and it contains some datasets on the legal domain. Moreover, files are structured in XML, which makes them easier to handle: for these reasons, TERMCAT files have been used in the project described in this dissertation.

\(^{16}\)http://tbx2rdf.lider-project.eu/converter/tbx2rdf.html
For the creation of Terminoteca RDF, Ontolex model was applied, as it features a module specially developed to represent terminologies. Since the datasets that are contained in Terminoteca RDF are multilingual, vartrans module, mentioned in Section 3.1, was a crucial part of the data modeling stage. The result is a repository of linked terminologies in some of the official languages of Spain that also provides translations for other languages of the European Union.

On the other hand, AGROVOC thesaurus is currently one of the biggest datasets in the Linguistic Linked Open Data cloud. Originally it was conceived as a printed vocabulary from the agricultural domain that contained terms in English, French and Spanish. In the year 2000 it was transformed into a relational database, and it has currently been published as Linked Data [28]. The RDF version of AGROVOC is represented in SKOS-XL, and it is composed of more than 35000 terms in 29 languages.

AGROVOC is linked with several datasets, such as GEMET for environment, TheSoz for social sciences or STW for economics. The property selected to link concepts between different dataset is skos:exactMatch, since it is not as restrictive as its formal equivalent in OWL owl:sameAs.

The conversion of AGROVOC into RDF was accomplished thanks to VocBench, a collaborative ontology and RDF editing tool, whose development was supported by the Food and Agriculture Organization of the United Nations (FAO). AGROVOC offers open access through several endpoints such as SPARQL, on-line search and RDF dumps\footnote{http://aims.fao.org/vest-registry/vocabularies/agrovoc}.

It is also worth mentioning the work carried out in the conversion of the bilingual dictionaries behind Apertium\footnote{https://www.apertium.org}, a free-open-source machine translation system, developed by the Spanish Government and several Spanish universities [29]. This contribution includes the conversion of these dictionaries into RDF, resulting in 22 linguistic datasets that are now part of the Linguistic Linked Open Data cloud. To do so, the lemon model was applied, as well as a specific module to represent translations that will be described in the last part of this document.

To sum up this section, the vocabularies used in each one of the works mentioned have been gathered in Table 3.

<table>
<thead>
<tr>
<th>Project</th>
<th>Vocabularies applied</th>
</tr>
</thead>
<tbody>
<tr>
<td>IATE/EMN</td>
<td>SKOS and Ontolex</td>
</tr>
<tr>
<td>Terminoteca RDF</td>
<td>Ontolex</td>
</tr>
<tr>
<td>AGROVOC</td>
<td>OWL and SKOS</td>
</tr>
<tr>
<td>Apertium</td>
<td>lemon</td>
</tr>
</tbody>
</table>

\textbf{Tab. 3: Project and vocabularies applied.}
3.2 Linguistic Linked Open Data Cloud

The overall purpose of RDF format and the Web of Data is that, starting from one data source, the user can get information from different sources connected by RDF links.

With this aim, Linked Open Data project has identified datasets available under open licenses, converted them into RDF format and linked them to give birth to the Linked Open Data cloud\(^\text{19}\) [5].

The Linked Open Data cloud, as the main source of Linked Data, has been deeply explored. It is divided in subclouds per domain, namely, Geography cloud, Government cloud, Media cloud, etc.; and each of them is composed by interlinked datasets belonging to that field. Accordingly, for this context, the Linguistic Linked Open Data cloud\(^\text{20}\) (LLOD) seems very convenient, since it provides exclusively linguistic resources sorted by typology:

- Corpora
- Terminology, thesauri and Knowledge Bases
- Lexicons and Dictionaries
- Linguistic Resource Metadata
- Linguistic Data Categories
- Typological Databases

Each one of the previously mentioned linguistic resources in RDF do belong to the LLOD as well: IATE, Terminesp, AGROVOC and GEMET. They are interlinked with other linguistic resources allowing a quick and easy access to information (see Figure 12).

\(^{19}\) http://lod-cloud.net/
\(^{20}\) http://linguistic-lod.org/llod-cloud
Fig. 12: Graphic representation of the Linguistic Linked Open Data Cloud.

Resources in the Linguistic Linked Open Data cloud are classified depending on their typology (previously listed). However, they have not been sorted by domain. For this reason, the result of this project will give birth to the first subcloud that gathers linguistic resources from a specific domain: the Linguistic Legal Linked Open Data cloud.
3.3 Legal Language Resources

3.3.1 Current Status

Nowadays, since society is living a huge trend of Open Data and public information, the number of documents generated online is becoming an issue, since accessibility and searchability can not cope with them. Also, the quality of these documents has been compromised: **unimportant and obsolete legal information is available for everyone.** This situation leads to the acquisition of incorrect knowledge and further misunderstandings with justice. For this reason, information retrieval (IR) and term extraction techniques are very necessary in this context. Some issues that IR tools need to take into consideration when applied to the legal domain [30]:

- Volume
- Document size
- Structure
- Heterogeneity of document types
- Self-contained documents
- Legal Hierarchy
- Temporal aspects
- Importance of citations
- Legal terminology
- Audience
- Personal data
- Multilingualism and multi-jurisdictionality
- Shortage of legal resources

There is a real need to improve the situation of legal localization: lexical resources and legal corpora may not be enough to find proper equivalents. Legal terminology may present issues such as the following [1]:

- “Terms of art”: this expression makes reference to those words that have a different meaning in the legal context than in an ordinary context. The expressions are often made up by law practitioners and they are born from the usage. Therefore, they do not have an official meaning and external users of law do not always interpret them in the right way.
Within a legal environment, these “terms of art” can also vary their meaning depending on the context (and jurisdiction). Polysemy represents a huge drawback when trying to understand an isolate term; for this reason, the meaning will always be linked to the text in which the terms is contained.

Sometimes law uses vague terms that allows more flexible rules to be changed in the future. It can be hard to find the right interpretation for these terms that often changes depending on the jurisdictions and laws applied to certain territories.

Also, the imprecision of the ordinary language is in the same way problematic in legal contexts. In some cases, it has been accepted to reinterpret given constructions depending on the situation, avoiding the literal meaning of the term.

One extra complication, not so language-related, are cross-references. Legislative texts are replete with cross-references which can difficult their readability and also point to irrelevant information for the user.

These difficulties are the main motivations why researchers are trying to structure, organise and link legal documents and linguistic assets. Apart from translators, multilingual open access to legal information is also required for judges, lawyers, legal drafters and scholars.

Furthermore, other decision makers such as enterprises, public administrations and citizens, subject to regulatory compliance would also be direct benefactors from these open and linked resources.

### 3.3.2 Legal Language Resources in RDF

Due to the hindrances mentioned in the previous section, legal language resources in the Web of Data are scarce. With the aim of solving these issues and improve the current status of Open Legal Data on the web, an experiment to link a term bank of multilingual and multijurisdictional legal data was performed [31]. The datasets taking part of this work include IATE, Creative Common licenses, documents from the World Intellectual Property Organization (WIPO) and other relevant resources such as DBpedia\(^{21}\) and Lexvo\(^{22}\).

To carry this process out, first they chose a collection of top concepts from the WIPO glossaries as well as from the copyright related glossaries. Afterwards, they mapped these top concepts to IATE. Finally, Creative Commons terms were added, including the different versions, jurisdictions and languages. The result was published as RDF, which means that it can be linked with other open legal resources. Some conclusions regarding the advantages of linking legal data were drawn afterwards:

\(^{21}\) [https://wiki.dbpedia.org/](https://wiki.dbpedia.org/)

\(^{22}\) [http://www.lexvo.org/](http://www.lexvo.org/)
Since data are structured in a formal model, a clear separation and term identification is generated.

Links between different entries allow an easy identification of a term in one language and its equivalent in several other languages.

Thus, localisation and translation processes within this domain will be heavily improved.

In the same manner, a structured navigation between general and specific terms of a given jurisdiction is provided.

Thanks to the metadata in which information about the sources of every document is contained, it is possible to perform comparative analysis of different resources.

Another very apposite project is Eunomos, described as an “advanced legal document and knowledge management system based on legislative XML and ontologies” [1]. Due to the technologies used, this project could be regarded as a counterpart of Lynx. It presents legal information in a structured and organised way, dividing laws by domain and keeping track of their updates.

It is intended to be multilingual and multilevel, which means that can contain multijurisdictional information and be used at EU level since it keeps separate ontologies for each system.

The system is based on the Legal Taxonomy Syllabus ontology, for terminology management of the European Directives, which helps with the extraction and modeling of legal concepts.

On the other hand, EuroVoc, the multilingual and multidisciplinary thesaurus that covers the activities of the European Union was originally published in XML-Eurovoc, a format that does not allow to import other thesauri published in different formats.

After much discussion and various proposals to decide the SKOS structure to be applied, currently, EuroVoc is available in SKOS and RDF [32]. It is also aligned with other relevant resources at European Level such as UNESCO and GEMET thesauri. It can be also accessed through a website search application and through a SPARQL endpoint developed by PoolParty. 23

The Publication Office of the European Union was, alongside some European universities, one of the supporters of the EuroVoc conversion. However, their work in the Semantic Web field does not end there. With the aim of publishing an accessible version of the publications and bibliographic resources of the European Union, they have developed the CELLAR repository, an information system based on semantic technologies that supports indexing, search and information retrieval tasks. Such documents, produced by the 28 member states of the European Union, were already available on-line thanks to the Eur-Lex portal. 24

23 https://lynx.poolparty.biz/PoolParty/sparql/Eurovoc4.3
24 https://eur-lex.europa.eu/
Still, many of the formats in which they were published are not machine processable nor structured (PDF, TXT); thus, their accessibility was limited. CELLAR resources, on the other hand, are semantically described by an ontology, which provides open access, long term preservation, indexing and retrieval services [33].

Lastly, another pioneer initiative in developing Semantic Web technologies in the legal domain was LOIS project, focused on Lexical Ontologies for Legal Information Sharing. In this project two types of information associated to legal texts are spotted: a conceptual model of the legal domain and a vocabulary to lexicalise such concepts. Accordingly, this model provides the necessary structure to build a semantic lexicon that contains the requirements of the language in the legal domain. It contains terminology from national and European legislation from the consumer law domain. Thus, LOIS acts as multilingual knowledge base that allows the user to perform a hierarchical search of concepts and their relations [34].

3.4 Linguistic Knowledge Graphs

As stated the introduction, the Linguistic Linked Open Data cloud gathers language resources in RDF published as Linked Data. Some of these resources are Linguistic Knowledge Graphs, structures that represent linguistic information through entities and relations.

The major resource contained in this graph is DBpedia, a vast network that structures data from Wikipedia and links them with other datasets available on the Web [35]. The result is published as Open Data available for the consumption of both humans and machines.

The other nucleus of the LOD Cloud is BabelNet, a large multilingual semantic network, generated automatically from various resources and integrates the lexicographical information of WordNet and the encyclopaedic knowledge of Wikipedia [36]. BabelNet also applies Machine Translation to get information from several languages. As a result, BabelNet is considered an encyclopaedic dictionary that contains concepts and named entities connected thanks to a great amount of semantic relations.

WordNet is one of the most known Linguistic Knowledge Graphs, since it is a large online lexical database that contains nouns, verbs, adjectives and adverbs in English [37]. These words are organised in sets of synonyms that represent concepts. WordNet uses these synonyms to represent word senses; thus, synonymy is WordNet’s most important relation. Four additional relations are also used by this network: antonymy (opposing-name), hyponymy (sub-name), meronymy (part-name), troponymy (manner-name) and entailment relations.

However, there are other semantic networks (considered linguistic knowledge graphs) that does not appear in the LOD Cloud but are also worth to mention.

This is the case of ConceptNet, a semantic network designed to represent common sense and support textual reasoning about documents in the real world. It represents part of human experiences and tries to share this common-sense knowledge with machines. ConceptNet is oft integrated with natural language processing
applications to speed up the enrichment of AI systems with common sense [38].

In addition, other knowledge bases have also been built based on the architecture of Wordnet. **EuroWordNet** was a project build by 8 different wordnets (nets of words) in English, Dutch, Italian, Spanish, French, German, Czech and Estonian. Each wordnet is considered as an “autonomous language specific ontology”[39]. The Italian variant, for instance, known as **ItalWordNet**, served as a foundation for a legal knowledge base, **JurWordNet** [40], a sound initiative to handle lexical polysemy of legal terms during the ontology building process.

### 3.5 Involved Technologies

Technology in the language field is applied to many different areas such as Natural Language Processing and Computational Linguistics. In this work, two main groups of language technologies have been studied and used: term extraction technologies and RDF and Linked Data technologies (applied to the linguistic domain). Aiming to choose the most appropriate for the purposes of this thesis, a review of the available technologies has been developed.

#### 3.5.1 Term extraction technologies

Originally, term extraction has been carried out by humans: domain experts and experienced terminologists. This is, probably, the best method to master this activity. However, it is also the most expensive and time-consuming. For this reason, computer scientists have been developing computer assisted term extraction applications based on statistical methods to handle texts in many different languages. Those term extraction applications based on linguistic methods are definitely more accurate, but they are also expensive; thus, only developed for the major languages.

Still, both terminology professionals and term extraction tools present many problems with disambiguation: synonymy and homonymy. Other examples of issues with automatic terminology extraction are listed below [41]:

Another difficult situation spotted by the developers of **TermSuite** comes up with parallel corpora, as this kind of documents are only available in the most widely spoken languages, and many other remain uncovered [42]. To solve this situation, TermSuite was developed as the first multilingual term extraction tool from comparable corpora (not parallel documents) to manage 7 languages from 5 different families.

By comparable corpora, they mean “sets of texts in different languages that are not translations of each other”. This innovative feature is the solution to the lack of corpora to be used by the tool. It is interesting as well to explore how the tool get to be multilingual: firstly it performs a monolingual term extraction and afterwards it carries out a bilingual term alignment. The stages of the architecture are divided as follows:

- Text pre-processing: includes text extraction, data cleaning, language recognizers, etc.
3.5 Involved Technologies

- Linguistic analysis: includes word tokenizers, part-of-speech taggers, syntactic parsers, etc.
- Term extraction: includes single-word terms (SWT), multiword terms (MWT) and related tasks.
- Term alignment: includes SWT and MWT alignment, and machine translation.

For instance, the developers of **TBXTools** used linguistic and statistical methods for multiterm extraction in specialised corpora [43]. They tried to create an alternative to the previous TE tools developed which were language dependent.

The methods they use work as follows:

- Statistical methods: they make use of n-grams and stop words. They also allow some normalizations such as capital letter normalization, morphological normalization and nested candidate detection.
- Linguistic methods: these methods include morphosyntactic pattern and a tagged corpus. Regular expressions and lemmatization techniques are as well used if required.

When it comes to corpus and corpora management, **Sketch Engine** is currently one of the leading technologies, used by a several types of professional fields [44]: lexicography, language teaching, translation, terminology, language technology companies and universities. This tool is very interesting within this project from the terminology extraction point of view.

Sketch Engine uses a grammar, lemmatisers, several corpora and statistic methods to analyse domain corpus uploaded by the user. They are able to provide comparisons against reference corpora available for 60 languages.

Several techniques can be applied to identify term candidates from a given text: comparing two corpora, measuring term frequency, tokenising, lemmatising, POS-tagging... Previously described tools already apply these methods independently. The innovation in Sketch Engine is that it combines all of them, creating an environment for everyone to easily find terms in a domain. Top term candidates will be the ones with the highest frequency in the domain corpus in comparison with the reference corpus.

### 3.5.2 Technologies for RDF modeling and linking

In this section, RDF management tools have been gathered, specially those that allow the conversion of various files into RDF. This step is crucial when working in the Semantic Web field, since the majority of the data generated nowadays are published in a heterogeneous variety of formats that needs to be homogenised.

Considering that independent RDF files by themselves lose the potential of this format, most of the tools reviewed also offer linking services. Moreover, some of them allow ontology and taxonomy management, and also term extraction services.
One of the most complete open source tools in this field is OpenRefine\(^{25}\) (formerly Google Refine), that allows cleaning and structuring data, converting them to other formats and linking the result with other resources in the web \[45\]. This service is published as a website application with a very intuitive interface that allows non-computer-scientist professionals making good use of it.

The tool supports several input formats such as CSV, XML and XLS, that are the most common methods of storing data. After the cleaning tasks, data can be structured in an RDF skeleton by using properties from several vocabularies and ontologies (SKOS, FOAF, OWL, etc); this skeleton can be exported afterwards to be directly applied to future files.

As per the linking features, with OpenRefine a file in RDF can be linked with other RDF files or with external knowledge bases. Its reconciliation service offers the two possibilities: on the one hand, the user can configure this service to link a local RDF file with a SPARQL endpoint (powered or not by Virtuoso) and on the other hand, the tool can create links between two RDF files.

Also, Virtuoso Sponger\(^{26}\) has been specially conceived for the conversion of files into RDF, generating Linked Data from many different source formats\(^{27}\). It is integrated with the Virtuoso SPARQL platform and it also supports many other serialization and representation formats apart from RDF.

Apart from these conversion tools, complementary technologies to manage RDF have also been developed. This is the case of RML, a generic language used to map relational databases into RDF \[46\]. In contrast to the previously mention techniques, RML is able to map several source files at once, regardless of the domain and format. This supposes an advantage over other tools that can only work with one type of source file, and also for data publisher, since RML prevents them from installing several tools depending on the data format.

On the other hand, VocBench\(^{28}\) is focused on thesauri management according to the SKOS and SKOS-XL standards \[47\]. It has been developed by the FAO\(^{29}\) and the University of Rome, and now it is being maintained by the Publications Office of the European Union, what means that it is presented as an open access application. The novelty of this application is its collaborative character: it allows to assign different user profiles depending on the needs of the thesauri. Some of the most common users are Ontology Editors, Terminologists, Validators and Publishers.

A similar service is provided by PoolParty\(^{30}\), a commercial website application developed by the Semantic Web Company for thesauri, taxonomies and ontologies management \[48\]. However, PoolParty can also be used to create links between data different data sources. This service is not open, but it is still relevant for the project since the developers are part of the Lynx consortium. As it was previously mentioned, Poolparty is specially interesting for ontology management, but a Eurovoc

\(^{25}\) http://openrefine.org/
\(^{26}\) http://vos.openlinksw.com/owiki/wiki/VOS/VirtSponger
\(^{27}\) http://vos.openlinksw.com/owiki/wiki/VOS/VirtSpongerCartridgeSupportedDataSourcesNonRDF
\(^{28}\) http://vocbench.uniroma2.it/
\(^{29}\) http://www.fao.org/home/en/
\(^{30}\) https://www.poolparty.biz/
3.5 Involved Technologies

SPARQL endpoint it also provided by this service, which means a very valuable asset for this project, since this thesaurus covers some subfields of the legal domain.

Another key linking service is **Silk**, the Linked Data Integration Framework\(^{31}\), an application that has been widely used by Semantic Web technologists over time to create links between entities [49]. The tool uses the SPARQL protocol to access data sources both locally and remotely. It is presented as a website application with a graphical editor that eases the conversion and linking tasks.

Last, but not least, **Karma**\(^{32}\) is a semi-automated website application developed by the University of Southern California for mapping structured data into the Semantic Web [50]. The difference between this service and other systems lies on the automation of the process: Karma allows to automatically create RDF files from a database while other tools do not. Moreover, the user can control the automatic process from an easy graphic user interface to fix, for instance, incorrectly assigned semantic types.

Although all of the tools mentioned are key applications in the Semantic Web field, OpenRefine was chosen to be involved in this project since it is published in an open format, as a website application with a simple and clear interface. The tool can be easily managed by users that are not specialised in the computer sciences area, while some of the tools above are intended for developers since they require coding skills for their installation and handling.

\(^{31}\) http://silkframework.org/

\(^{32}\) http://usc-isi-i2.github.io/karma/
3.6 Current needs

In the light of this state of the art, several needs have been identified:

- Many of the available language resources nowadays are not in machine readable formats yet. General resources on the Web are usually published in PDF, fact that generates difficulties when handling them. Moreover, many legal language resources are still being published in a physical format.

- In the same way, the existence of legal language resources on the Internet is scarce. Specifically, there is a lack of resources in the three subdomains involved in Lynx project: Labour Law, Data Protection and Industrial Standards. This problem is generated by the complexity of legal terminology. Legal terms are not easy to define, since their meaning changes every time law is modified.

- Lately, governments and institutions, specially in the European Union, have been publishing their legislation on the Internet. However, most have not adopted neither an open format nor linking standards yet: the greatest part of legal information is published as PDF files in private applications. There is a need to interlink those legal documents, providing quick access to multilingual and multijurisdictional legal information.

- Linking formats (RDF) are not very extended either. The lack of legal documents published as RDF is preventing the free access to information in the legal domain. More legal language resources (like those identified in Section 4.1) need to be converted into RDF.

- Documentation process is a key task that represents a great part of the scientific work. Terminological projects also involve documenting the features of each resource that has been generated. However, this stage tends to be forgotten. This is why there are not many language resources whose documentation is open; therefore, the documentation of language resources in the legal domain should be also an issue to amend.

- Finally, since there are not many resources in RDF and their documentation is not correctly published, there are a lot of datasets missing in the Linked Open Data cloud. A pending task is to search and identify relevant resources to be linked to the LLOD so they can share information.

This is one of the goals of the Legal Knowledge Graph. Therefore, in order to build it, a Linguistic Legal Linked Open Data cloud needs to be generated first to support the Legal Knowledge Graph with linguistic information.
4 Contribution

As stated in the previous section, the goal of this thesis is to contribute to the Linguistic Linked Open Data cloud with resources of the legal domain. By increasing the representativeness of legal resources in the LLOD, we will also contribute to initiatives such as the one proposed in the Lynx project of creating a Legal Knowledge Graph in which documents (law, case law, doctrines, literature...), metadata, ontologies, standards, and so on, are interconnected to facilitate the access to legal data. With this aim, the tasks performed and described in this thesis have been represented as a workflow to guide the reader (see Figure 13), and are specified below:

- **Identification of existing language resources to be reused**: search and evaluation of linguistic resources in the legal domain that are available and can be reused in this project. This stage also comprises those language resources that do not belong to the legal domain but can be useful to get more information about legal terms.

- **Creation of new of linguistic resources**: thanks to the legal corpora provided by Lynx partners, terminological glossaries from the legal domain have been created. This stage includes analysing and testing of a set of term extraction tools.

- **Conversion of resources into RDF**: in this stage, several linguistic assets gathered both in the identification and in the creation stage are converted into RDF. For this purpose, so-called RDF templates or skeletons are created with the OpenRefine tool.

- **Linking with other existing resources in the Linguistic Linked Open Data cloud (LLOD)**: the glossaries resulting from the previous stages are linked to larger knowledge bases that are already part of the LLOD, with the aim of having the glossaries handled in this project also in this cloud.

If the datasets require it, two subprocesses are applied:

- Transformation of the resource into a machine readable format (if a dataset is published in non structured or non machine readable formats)

- Adaptation of the RDF skeleton (in the case that a dataset is already in RDF but its schema is different to the one developed here)

Figure 13 shows the complete diagram with the tasks and sub-tasks involved in this work.
Fig. 13: Flow diagram showing the stages of the process.
The methodology followed is based on a sound methodology to publish data on the Web of Data defined in [51], which identifies the following processes:

- Data exploration
- URI naming strategy
- Data modeling
- RDF generation
- Linking

Some peculiarities and specificities can be found in the legal domain. For this reason, several steps in this methodology have been slightly adapted as required by the legal domain.

To start with, a previous process is required prior to the data modelling, because most of the resources identified are published in outdated formats. Therefore, a previous step in which resources had to be transformed to machine-readable formats is necessary. This is mainly due to the fact that this domain has traditionally been reluctant to technologies in general.

The other step that is a bit underspecified and would need some domain adaptation is the linking step. Since the converted resources in this domain are commonly bound to one jurisdiction, the linking task is a highly complex one. In this sense, some further recommendations or suggestions could be added to guide users following this methodology when converting resources of the legal domain (see Section 4.4).

4.1 Identification of existing resources

The first step to develop a Linguistic Legal Knowledge Graph is to discover if there are already existing resources that can be reused in this project.

In order to identify relevant resources, three different paths were explored:

- General web search
- Lookup of resources described in papers from the specialized literature
- Search in data portals specialized in language resources

The latter includes ELRC-SHARE repository\textsuperscript{33} (used for documenting linguistic resources by the European Language Resource Coordination), Retele\textsuperscript{34} (network of excellence for language technologies in Spain), CLARIN\textsuperscript{35} (European research

---

\textsuperscript{33} https://www.elrc-share.eu/
\textsuperscript{34} http://catalogo.retele.linkeddata.es/
\textsuperscript{35} https://www.clarin.eu/
infrastructure for language resources) and the OLAC Language Resource Catalog\(^{36}\) (unified portal for language resource search).

In the same way, a classification by domain has also been made: legal domain related resources have been marked as preferred. However, some general terminologies were also taken into account, since they own other important features, useful for the project.

Due to the huge amount of information and open data available nowadays, it is essential to establish these limits to gather only the relevant resources. In the case that more types of datasets are required, they will be harvested at a later stage.

The structure followed in this section has been chosen according to the relevancy of each resource for this specific project. Therefore, the first block of resources are those already published as Linked Data. Then, since this work is part of a European project, the second block is composed by resources supported or developed by the European Union. Consequently, the third group is formed by assets supported or developed by the United Nations, as it is a consolidated organisation with great significance in the legal field. The next two sets contain resources developed by other companies and organisations, but they are sorted by format: the first one gathers resources in machine readable formats and the last one collects documents in PDF. Finally, the last section shows legal language resources that have been archived.

### Resources in the Linguistic Linked Open Data cloud

Thus, some of the resources already linked to the LLOD cloud that have been regarded as valuable assets are:

- **STW Thesaurus** for Economics: a thesaurus that provides a vocabulary on any economic subject. It also contains terms used in law, sociology and politics (monolingual, en).

- Copyright Termbank: a multi-lingual term bank of copyright-related terms has been published connecting WIPO definitions, IATE terms and definitions from Creative Commons licenses (multilingual).

- **EuroVoc**: a multilingual and multidisciplinary thesaurus covering the activities of the EU. It is not specifically legal, but it contains pertinent information about the EU and their politics and law (multilingual).

- **AGROVOC**: a controlled vocabulary covering all the fields of the Food and Agriculture Organization (FAO) of the United Nations. It contains general information and it has been selected since it shares many structures with other important resources (multilingual).

- **IATE RDF**: a terminological database developed by the European Union which is constantly being updated by translators and terminologists. Amongst

other domains, the terms are related with law and EU governments (multilingual).

All resources published by the European Union are, clearly, a priority. These mentioned above are already transformed into RDF.

However, those published in other formats are to be considered as well. Machine readable formats, such as TBX, CSV and XLS are preferred. Exceptionally, resources published in non-machine-readable formats could be considered, depending on the relevancy of their content.

**Resources supported by the European Union**

Consequently, the following resources published by the European Union have also been listed as usable, although they are not included in the LLOD:

- **EUR-Lex**: a corpus containing EU public documents such as directives, legislative proposals, reports, judgements, international agreements, etc. These documents are presented in HTML and PDF, but there is also a SPARQL Endpoint to search for RDF structures (multilingual).

- **European Data Portal**: a site that harvests the metadata of Public Sector Information available on public data portals across European countries. It is also available in the European languages and provide the files in several formats, including RDF, CSV and XML (multilingual).

- **INSPIRE Glossary**: a term base developed by the INSPIRE Knowledge Base of the European Union. Although this project is related with the field of spatial information, the glossary contains general terms and definitions that specify the common terminology used in the INSPIRE Directive and in the INSPIRE Implementing Regulations (monolingual, en).

- **EUGO Glossary**: a term base addressed to companies and entrepreneurs that need to comply with administrative or professional requirements to perform a remunerated economic activity in Spain. This glossary is part of a European project and contains terms about regulations that are valuable for Lynx purpose (monolingual, es).

- **GEMET**: a general thesaurus, conceived to define a common general language to serve as the core of general terminology for the environment. This glossary is available in RDF and it shares terms and structures with EuroVoc (multilingual).

- **TermCoord**: a portal supported by the European Union that contains glossaries developed by the different institutions. These glossaries cover several fields including law, international relations and government (multilingual).

Although the resources are available in PDF, at some point these documents could be treated and transformed into RDF if necessary.
Resources supported by the United Nations

In the same way, the United Nations also counts with consolidated terminological resources. Given their intergovernmental domain, the following resources have been selected:

- **UNESCO Thesaurus**: a controlled list of terms intended for the subject analysis of texts and document retrieval. The thesaurus contains terms on several domains such as education, politics, culture and social sciences. It has been published as a SKOS thesaurus and can be accessed through a SPARQL endpoint (multilingual).

- **InforMEA Glossary**: a term bank developed by the United Nations and supported by the European Union with the aim of gathering terms on Environmental Law and Agreements. It is available as RDF and it will be upgraded to a thesaurus during the following months (multilingual).

- **International Monetary Fund Glossary**: a terminology list containing terms on economics and public finances related with the European Union. It is available as a PDF downloadable file; however, it may be transformed as a future work (multilingual).

Relevant resources in machine-readable format

On the other hand, other linguistic resources (not supported by the EU nor the UN) have been spotted. Some of them are already converted into RDF:

- **TERMCAT Glossaries**: a terminological database supported by the government of Catalonia. It contains general equivalences in several languages. Part of these terms have been converted into RDF and they are part of the Terminoteca RDF Project. They can be accessed through a SPARQL endpoint (multilingual).

- **German Labour Law Thesaurus**: a thesaurus that covers all main areas of labour law, such as the roles of employee and employer; legal aspects around labour contracts. It is available through a SPARQL endpoint and as RDF downloadable files (monolingual, de).

- **JuriVoc**: a juridical thesaurus developed by the Federal Supreme Court of Switzerland in cooperation with Swiss legal libraries. It contains juridical terms arranged in a mono-hierarchic structure (multilingual).

- **Library of Congress**: a system that makes documents from the Library of Congress available as a public domain datasets. It is composed of documents, books and agreements published as RDF, JSON and other structured formats (monolingual, en).
4.1 Identification of existing resources

- **SAIJ Thesaurus**: a thesaurus that organises legal knowledge through a list of controlled terms which represent concepts. It is available in RDF and intended to ease users’ access information related to the Argentine legal system that can be found in a file or in a documentation centre (monolingual, es).

- **CaLaThe**: a thesaurus for the domain of cadaster and land administration that provides a controlled vocabulary. It is interesting because it shares structures and terms with AGROVOC and the GEMET thesaurus, and it can be downloaded as an RDF file (monolingual, en).

- **CDISC Glossary**: a glossary contains definitions of terms and abbreviations that can be relevant for medical laws and agreements It is available in several formats, including OWL (monolingual, en).

**Relevant resources in PDF**

Finally, one last resource available in other PDF has also being considered due to different facts:

- **Connecticut Glossary**: a glossary that contains legal terms published by the Judicial Branch of the State of Connecticut. It can be transformed into a machine-readable format and from there into RDF since it provides with equivalences of legal terms from English into Spanish (bilingual).

**Archived Resources**

During this process, additional relevant resources have also been identified. However, these resources have been archived for various reasons: they are published in formats that can’t be handle, they are not available or they are not supported any longer by their publishers.

<table>
<thead>
<tr>
<th>Name</th>
<th>URI</th>
<th>Availability</th>
<th>Domain</th>
</tr>
</thead>
<tbody>
<tr>
<td>CFR SKOS Vocabulary</td>
<td>-</td>
<td>Not available</td>
<td>Legal</td>
</tr>
<tr>
<td>Legal RDF Vocabulary</td>
<td></td>
<td>Not available</td>
<td>Legal</td>
</tr>
<tr>
<td>LexMIL</td>
<td><a href="http://www.uv.es/lexmil/">http://www.uv.es/lexmil/</a></td>
<td>Not available</td>
<td>Legal</td>
</tr>
<tr>
<td>RDF dictionary</td>
<td><a href="http://rdfs.vocabulary.sourceforge.net/">http://rdfs.vocabulary.sourceforge.net/</a></td>
<td>Not available</td>
<td>Legal</td>
</tr>
<tr>
<td>Lexdata</td>
<td><a href="http://www.lexdata.org/">http://www.lexdata.org/</a></td>
<td>Not available</td>
<td>Legal</td>
</tr>
<tr>
<td>Law</td>
<td><a href="http://lawin.org/legal-rdf/">http://lawin.org/legal-rdf/</a></td>
<td>Available</td>
<td>Legal</td>
</tr>
<tr>
<td>UKSCC</td>
<td></td>
<td>Not available</td>
<td>Legal</td>
</tr>
<tr>
<td>HTTR</td>
<td><a href="http://www.attig.cnr.it/BancheDataAndrade/IndiceEng.php">http://www.attig.cnr.it/BancheDataAndrade/IndiceEng.php</a></td>
<td>Available</td>
<td>Legal</td>
</tr>
<tr>
<td>EuroWordNet</td>
<td><a href="http://projects.del.uva.nl/EuroWordNet/">http://projects.del.uva.nl/EuroWordNet/</a></td>
<td>Not available</td>
<td>Legal</td>
</tr>
<tr>
<td>termbases.eu</td>
<td><a href="https://www.termbases.eu/">https://www.termbases.eu/</a></td>
<td>Available</td>
<td>Diverse</td>
</tr>
<tr>
<td>UNHCR</td>
<td><a href="http://www.refugeethesaurus.org/hms/search.php?action=searchpage&amp;expand=1">http://www.refugeethesaurus.org/hms/search.php?action=searchpage&amp;expand=1</a></td>
<td>Available</td>
<td>Social</td>
</tr>
<tr>
<td>Webster Law Dictionary</td>
<td><a href="http://law.yourdictionary.com/">http://law.yourdictionary.com/</a></td>
<td>Available</td>
<td>Legal</td>
</tr>
<tr>
<td>Black Law Dictionary</td>
<td><a href="http://thesaurusdictionary.org/">http://thesaurusdictionary.org/</a></td>
<td>Available</td>
<td>Legal</td>
</tr>
<tr>
<td>Dudario jurídico ONU</td>
<td><a href="https://onutraduccion.wordpress.com/pref/dudario-juridico/#H">https://onutraduccion.wordpress.com/pref/dudario-juridico/#H</a></td>
<td>Available</td>
<td>Legal</td>
</tr>
</tbody>
</table>

Tab. 4: Archived resources.
Summary

To put all the pieces together, only available resources have been considered as reusable, and they are exposed in Table 5, along with a summary of their most remarkable features.

<table>
<thead>
<tr>
<th>ID</th>
<th>Name</th>
<th>Description</th>
<th>Language</th>
</tr>
</thead>
<tbody>
<tr>
<td>iate</td>
<td>IATE</td>
<td>EU terminological database</td>
<td>EU languages</td>
</tr>
<tr>
<td>eurovoc</td>
<td>Eurovoc</td>
<td>EU multilingual thesaurus</td>
<td>EU languages</td>
</tr>
<tr>
<td>eur-lex</td>
<td>EUR-Lex</td>
<td>EU legal corpora portal</td>
<td>EU languages</td>
</tr>
<tr>
<td>connect-legal-glossary</td>
<td>Connect Legal Glossary</td>
<td>Bilingual legal glossary</td>
<td>en, cs</td>
</tr>
<tr>
<td>unesco-thesaurus</td>
<td>UNESCO Thesaurus</td>
<td>Multilingual multidisciplinary thesaurus</td>
<td>en, es, fr, ru</td>
</tr>
<tr>
<td>lisa-congress</td>
<td>Library of Congress</td>
<td>Legal corpora portal</td>
<td>en</td>
</tr>
<tr>
<td>imf</td>
<td>International Monetary Fund</td>
<td>Economic multilingual terminology</td>
<td>en, de, es</td>
</tr>
<tr>
<td>eu-go-glossary</td>
<td>EUGO Glossary</td>
<td>Business monolingual dictionary</td>
<td>es</td>
</tr>
<tr>
<td>cdisc-glossary</td>
<td>CDISC Glossary</td>
<td>Clinical monolingual</td>
<td>en</td>
</tr>
<tr>
<td>stw</td>
<td>STW Thesaurus for Economics</td>
<td>Economic monolingual thesaurus</td>
<td>en</td>
</tr>
<tr>
<td>esp</td>
<td>European Data Portal</td>
<td>EU datasets</td>
<td>EU languages</td>
</tr>
<tr>
<td>inspire</td>
<td>INSPIRE Glossary (EU)</td>
<td>General terms and definitions in English</td>
<td>en</td>
</tr>
<tr>
<td>saj</td>
<td>SALT Thesaurus</td>
<td>Controlled list of legal terms</td>
<td>es</td>
</tr>
<tr>
<td>calathe</td>
<td>CaLaThe</td>
<td>Cadastre vocabulary</td>
<td>en</td>
</tr>
<tr>
<td>gmet</td>
<td>GEMET</td>
<td>General multilingual thesaurus</td>
<td>en, de, es, it</td>
</tr>
<tr>
<td>informea</td>
<td>InforMEA Glossary (UNESCO)</td>
<td>Monolingual glossary on environmental law</td>
<td>en</td>
</tr>
<tr>
<td>copyright-termbank</td>
<td>Copyright Termbank</td>
<td>Multi-lingual term bank of copyright-related terms</td>
<td>en, es, fr, pt</td>
</tr>
<tr>
<td>glink</td>
<td>German labour law thesaurus</td>
<td>Thesaurus with labour law terms</td>
<td>de</td>
</tr>
<tr>
<td>jurivoc</td>
<td>Jurivoc</td>
<td>Juridical terms from Switzerland</td>
<td>de, it, fr</td>
</tr>
<tr>
<td>TERMCAT</td>
<td>TERMCAT</td>
<td>Terms from several fields including law</td>
<td>ca, en, es, de, fr, it</td>
</tr>
<tr>
<td>termcoord</td>
<td>Termcoord</td>
<td>Glossaries from EU institutions and bodies</td>
<td>EU languages</td>
</tr>
<tr>
<td>agrovoc</td>
<td>Agrovoc</td>
<td>Controlled general vocabulary</td>
<td>29 languages</td>
</tr>
</tbody>
</table>

Tab. 5: Set of available language resources identified.

The set of the harvested linguistic resources has also been visually represented in a graph, in which each dataset is coloured as per the domain covered (Figure 14.). A second version of the graph has also been created in order to make a distinction between those datasets in RDF and those in different formats (Figure 15.). The graphs also represent the relations between each asset, since most of those in RDF share structures and terms.
4.1 Identification of existing resources

Fig. 14: Datasets represented by domain.

Fig. 15: Datasets represented by format.
After this research process, several conclusions on the status of legal language resources can be drawn:

- Legal language resources are often generated for one single sub-domain of law. For this reason, it is more complex to find equivalences between terms in glossaries since they do not belong exactly to the same domain (within the law area).
- Usually the resources have been created only for the most common languages in Europe. Thus, minor languages are not represented.
- Datasets are heterogeneous. They present many different features including:
  - Format
  - Language
  - Lexical information
- Many of the resources gathered are supported by European Institutions. However, other interesting resources developed by old European projects have been disregarded since they are not maintained any longer.
- Conversely, other resources have been generated by small organisations or individuals, usually in obsolete formats that are no longer available or useful.

To perform the conversion and linking experiments in order to generate the first draft of the *Linguistic Legal Linked Open Data cloud*, two TERMCAT terminological glossaries have been selected. TERMCAT repository contains glossaries in several languages from many different areas such as agriculture, arts, sports, or economics, to mention but a few. The selected glossaries belong to the law domain, more specifically to the subdomain of collective negotiation, that belongs to the labour law area, and they contain terms in English and Spanish.

The glossaries are published in XML format, which is an advantage since they can be directly converted into RDF without any intermediate conversion process.

Many other datasets from the collection identified are also relevant for the purposes of this work. However, they will be handled as future work because they require more treatment: some of them are difficult to represent in RDF, others are published in PDF and need to be converted, etc. On the other hand, some other datasets contain more general information and they are published in languages that are not covered by the Lynx project, so linking them is not a priority.
4.2 Creation of new resources

4.2.1 Term extraction stage

Due to the shortage of linguistic legal resources for certain sub-domains, the creation of new datasets became a requirement. For this purpose, it was decided to perform a term identification task from corpora. The corpora were provided by the industrial partners representing the three pilots to be developed in the Lynx project, namely, the Spanish law firm CuatreCasas, the Austrian legal tech company Openlaws, and the Norwegian industrial certification company DNV GL (see website of the Lynx project\textsuperscript{37} for more details on the partners).

The specific sub-domains for which corpora was provided are spelled out below:

- CuatreCasas provided with a set of documents in Spanish composed by collective agreements. Thus, this pilot is related to the Labour Law domain.

- OpenLaws provided with a set of documents in English composed by regulations for data protection in the European Union. Thus, this pilot is related to the Data Protection domain.

- DNV provided with a set of documents in English composed by decisions and agreements in the maritime industry. Thus, this pilot is related to the Industrial Standards domain.

All these documents are published in a non-machine-readable format: PDF. To treat them, there were two options: either convert them into plain text (TXT) or find a term extraction tool that can work with PDF files.

Finding such tool has been one remarkable task in this project. There has been an extended process of search and evaluation of several open term extraction tools with the aim of selecting the most complete and suitable for this project.

Evaluated tools are listed as follows:

- Translated.net LABS Terminology Extraction system
- VocabGrabber
- TermoStat Web
- Sketch Engine
- Fivefilters Term Extraction system
- Termine
- Pootle Terminology Exraction System
- TBXTools

\textsuperscript{37}http://lynx-project.eu/
• TermSuite

For the evaluation, many different factors were considered, and several term extraction experiments were carried out.

Table 14 contains an overview of the main features of each tool. Some of them are the tool format, type of access, language filter and services provided, to mention but a few. It can be consulted at the end of this document, in the Annexes section.

Based on the results of the experiments performed to evaluate the tools, some general issues have been spotted:

• It has been proved that a great part of the words identified tested extraction tools are not really terms, and the term lists generated must be postedited by a terminologist to ensure the quality of the extracted terms.

• Normally, term extraction tools do not give any other terminological information apart from the list of terms. A terminologist is also required at this point to look for synonyms, grammatical information, relations between terms, usage examples, etc.

• Term identification can be problematic when a complex syntax is applied (especially multi-word terms). As stated above, in order to improve the results, linguistic knowledge needs to be applied, but it is expensive.

Some of the tools, such as VocabGrabber and Five Filters are website applications that do not allow to process input PDF files: the user needs to copy plain text in the website interface. This is a major drawback when processing large corpora. On the other hand, Termine does not apply a language filter, which means that the tool is not able to detect the language of the input document, and the candidate terms are less accurate.

After several tests with the nine extraction tools selected, the results were manually evaluated and it was decided to use Sketch Engine because the quality of the results was higher than the ones provided by the rest of the tools.

Sketch Engine also offers additional features that are very useful for the purposes of the project, such as corpora management and term linking with external resources [52]. Other decisive features for this choice are the following:

• It can filter the input files by language, providing better results.

• It can extract both single and compound terms (other tools only extract single terms).

• It does not extract stop words (other tools do extract them).

• It allows the creation of a corpus of several files (other tools only accept the extraction from one single file).

• If also provides several selectable reference corpora, depending on the interests of the user.
4.2 Creation of new resources

- It accepts several input formats, including PDF.
- The output is available both in CSV and TBX format.
- Although this tool is available under payment, it offers a 30-days free trial.

The tool returns two term lists per set of documents analysed: **a list composed by single terms and a list containing compound terms.** The lists include candidate terms and an indicator of the frequency of each term in the given corpus (Figure 16). Each raw list has 100 terms (200 in total) that are to be evaluated to determine which terms have been correctly extracted, which are relevant for the domain, etc.

![Fig. 16: Example of candidate terms in Sketch Engine.](image)

The result of this “cleaning stage” is compounded by **three glossaries (one per pilot)** that have been formatted and organised in XLSX files. Terminological glossaries contain around one hundred terms each: **50 of them are single terms and the remaining 50 terms are compound.**

In the glossaries, there are eight columns per entry describing each term. These attributes will be explained in Section 4.3 (Conversion into RDF), since these columns have been created according the properties chosen to represent the glossaries in RDF.

The decision to determine whether the candidate words proposed by the tool were real terms or not was helped by terminological tools such as IATE, that has already been described, and **Linguee**. The latter provides with term definitions and examples of the context where the searched terms are used. Many of these usage contexts are extracted from Eur-Lex, so they are considered reliable sources (example in Figure 17). The tool is also linked to Wikipedia, so information of this repository can also be consulted.

---

38 https://www.linguee.es
4.2.2 Term evaluation stage

The extracted terms by Sketch Engine have been evaluated by two linguists to reject those candidates that can’t be considered as terms (namely, general words or stop words), and also groups of letters that are not neither terms nor words (Figure 18).

The performance of the reference corpora used by the extraction tool to identify the terms in the input corpora has been also analysed. More specifically, two corpora have been chosen: a corpora composed by general documents and EUR-Lex, corpora from the legal domain.

For this experiment, corpora from the Industrial Standards domain has been used. This specific dataset has been chosen since it mostly gathers technical terms from maritime industry, and it was interesting to see the performance of a legal reference corpus against an industry input corpus.

As Table 6 shows, the use of one corpus or another does not make a big difference.

<table>
<thead>
<tr>
<th>Reference corpora</th>
<th>Legal terms</th>
<th>Neutral terms</th>
<th>Wrong terms</th>
<th>Industrial/Maritime terms</th>
<th>Total Terms</th>
</tr>
</thead>
<tbody>
<tr>
<td>General</td>
<td>3</td>
<td>41</td>
<td>77</td>
<td>79</td>
<td>200</td>
</tr>
<tr>
<td>Eur-Lex</td>
<td>1</td>
<td>34</td>
<td>80</td>
<td>85</td>
<td>200</td>
</tr>
</tbody>
</table>

Tab. 6: Corpora comparison.

What is relevant from these results is that, from 200 extracted terms, around 80 are identified as wrong terms. Figure 18 shows an excerpt of the extracted terms list, with examples of what has been considered as incorrect terms (in red), a term from the maritime industry field (grey cells), a general term (yellow cells).
Although the comparison of the results by using the two different reference corpora has been only performed applied to the Industrial Standards domain, for the two remaining domains a list of 200 terms was also generated and half of those terms were marked as incorrect. For this reason, the resulting glossaries that have been handled along this project contain around 100 terms.

4.3 Conversion into RDF

Once the existing relevant datasets have been identified and classified, and the three glossaries from Lynx corpora have been created, they pass to the RDF conversion stage.

The following five glossaries will be converted and linked:

- Cuatrecasas terminological glossary from the Labour Law domain in Spanish (New)
- Openlaws terminological glossary from the Data Protection domain in English (New)
- DNV terminological glossary from the Industrial Standards domain in English (New)
- TERMCAT glossary of Labour Law domain in English (Reused)
- TERMCAT glossary of Labour Law domain in Spanish (Reused)

These resources, when linked to other resources in the Linked Open Data cloud will give birth to the alpha version of a Linguistic Linked Open Data cloud in the Legal Domain.

Therefore, the first task is to organise and structure the files, namely, cleaning them to avoid introducing noisy data in the graph.

This process has been performed thanks to the GREL expressions in OpenRefine\(^39\) and the functions of Excel. Some “cleaning activities” are related with

\(^{39}\)https://github.com/OpenRefine/OpenRefine/wiki/GREL-Functions
consistency of the entries, for instance, removing duplicates or finding empty cells. However, the greatest part of this job has to do with the URI naming strategy that is described below.

### 4.3.1 URI naming strategy

The major part of the tasks mentioned in the previous section that involves the usage of **GREL expressions** is related to the creation of the URI of each entry: since the URI is based on the term itself, it has been necessary to remove spaces in compound terms, accents and special characters; and also the addition of other information.

The URI naming strategy adopted is based on the previous work in transforming TERMCAT files [26], that is at the same time based on the approach followed in the conversion of the Apertium Bilingual Dictionaries [29].

The idea is to connect the datasets with the existing Linguistic Linked Open Data cloud, as a contribution to the Terminoteca RDF project, being part of the Lynx project at the same time. Thus, the base URI appears as follows:


The identifier of each term will be composed by the written representation of the term in the corresponding language, along with its part of speech and the ISO 639-1 language code (e.g.:law-n-en).

Thus, for each entry, the naming strategy looks as the following examples:

<table>
<thead>
<tr>
<th>Term</th>
<th>Term URI</th>
</tr>
</thead>
<tbody>
<tr>
<td>Derogation</td>
<td><a href="http://linguistic.linkeddata.es/terminoteca/lynx/derogation-n-en">http://linguistic.linkeddata.es/terminoteca/lynx/derogation-n-en</a></td>
</tr>
<tr>
<td>Infringe</td>
<td><a href="http://linguistic.linkeddata.es/terminoteca/lynx/infringe-v-en">http://linguistic.linkeddata.es/terminoteca/lynx/infringe-v-en</a></td>
</tr>
<tr>
<td>Licitador</td>
<td><a href="http://linguistic.linkeddata.es/terminoteca/lynx/licitador-n-es">http://linguistic.linkeddata.es/terminoteca/lynx/licitador-n-es</a></td>
</tr>
<tr>
<td>Despido procedente</td>
<td><a href="http://linguistic.linkeddata.es/terminoteca/lynx/despido-procedente-n-es">http://linguistic.linkeddata.es/terminoteca/lynx/despido-procedente-n-es</a></td>
</tr>
</tbody>
</table>

Tab. 7: Examples of term URIs.

### 4.3.2 Modelling

Once the URI is created, the properties that are going to model the RDF skeleton need to be chosen based on the information that will be represented. It is convenient to make these decisions prior to the creation of the source file (CSV, XLSX, etc.) that is to be converted, in order to distribute the content accordingly.

Taking into account the current state of the glossaries (term lists), their objective (annotate legal documents) and their content (terms and some definitions), SKOS seems the most appropriate model to be applied to this project due to its simplicity and intuitive representation. These initial glossaries will evolve, gathering more information, and properties of the other models previously mentioned will be used.

By now, bearing in mind the information exposed in Section 3.1, the following SKOS properties have been selected to model the data contained in the glossaries that are to be converted into RDF:
4.3 Conversion into RDF

| skos:prefLabel | It assigns a preferred lexical label to a resource. Since RDF literals are defined as character strings with optional language tags, SKOS allows to assign multilingual labels. |
| skos:altLabel | It assigns an alternative lexical label to a concept. This is property assigns a different label from the preferred one and it is normally used to represent synonyms. |
| skos:definition | It shows a complete description of the meaning of the term. |
| skos:note | It shows a complete description of the meaning of the term. |
| skos:broad | It is used to assign a more general concept to the main entry. Consequently, skos:narrower expresses that one concept is more specific than another. |
| skos:topConcept | It links a concept scheme with the most general concepts contained in it (i.e. “animal”, “person”, “vehicle”). |
| skos:hasTopConcept | It is used to link a concept scheme to the top SKOS concepts. |
| skos:closeMatch | It links two similar terms that can be interchanged. (*This property is applied in the linking step, not in the initial RDF skeleton.) |

Tab. 8: SKOS properties applied.

To represent the metadata of each glossary in RDF, the DublinCore ontology was applied:

| dc:creator | This property is applied to show the creator of a resource. |
| dc:date | This property is applied to show the creation date of a resource. |
| dc:title | This property is applied to show the title of the resource. |
| dc:description | This property offers a brief description of the information contained in the resource. |

Tab. 9: DublinCore properties.
Thereby, before importing the glossaries in OpenRefine, they have been structured according to these SKOS properties, with the aim of easing the conversion process.

The following screenshots represent an example of this organisation:

<table>
<thead>
<tr>
<th>conceptid</th>
<th>prefLabel</th>
<th>definition</th>
</tr>
</thead>
<tbody>
<tr>
<td>controller-n-en</td>
<td>controller</td>
<td>The chief accounting officer of a business enterprise or an institution (such as a college)</td>
</tr>
<tr>
<td>derogation-n-en</td>
<td>derogation</td>
<td>Provision in an EU legislative measure which allows for all or part of the legal measure to be applied differently, or not at all, to individuals, groups or organisations</td>
</tr>
<tr>
<td>member-n-en</td>
<td>member</td>
<td>A person or legal entity having an interest as a partner or shareholder in a firm or other enterprise</td>
</tr>
</tbody>
</table>

Fig. 19: Example of the glossary structure (I).

<table>
<thead>
<tr>
<th>altLabel</th>
<th>altLabel</th>
<th>broader</th>
<th>topConcept</th>
<th>note</th>
</tr>
</thead>
<tbody>
<tr>
<td>annulment</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>E.g. &quot;Spain is member of the European Union&quot;</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Fig. 20: Example of the glossary structure (II).

As a result, after importing the glossary into OpenRefine and assigning the properties with the corresponding columns in the glossary, the following RDF skeleton is generated:

Fig. 21: Example of RDF skeleton in OpenRefine (I).
4.3 Conversion into RDF

The RDF output of this conversion presents the following structure:

```
<rdf:Description rdf:about="http://linguistic.linkeddata.es/terminoteca/lynx">
  <rdfs:label xml:lang="en">Industrial Standards Glossary</rdfs:label>
  <dc:creator rdf:resource="http://www.w3.org/2004/02/skos/core#ConceptScheme"/>
  <dc:created rdf:datatype="http://www.w3.org/2001/XMLSchema#dateTime">2018-06-20</dc:created>
  <dc:title xml:lang="en">Industrial Standards Glossary</dc:title>
  <skos:prefLabel xml:lang="en">Industrial Standards Glossary</skos:prefLabel>
  <skos:definition xml:lang="en">Glossary with terms extracted from legal corpora containing files with information about industrial standards.</skos:definition>
  <skos:hasTopConcept rdf:resource="http://linguistic.linkeddata.es/terminoteca/lynx/stiffener-n-en"/>
</rdf:Description>
```

```
<rdf:Description rdf:about="http://linguistic.linkeddata.es/terminoteca/lynx/stiffener-n-en">
  <skos:prefLabel xml:lang="en">stiffener</skos:prefLabel>
</rdf:Description>
```

```
<rdf:Description rdf:about="http://linguistic.linkeddata.es/terminoteca/lynx/girder-n-en">
  <skos:prefLabel xml:lang="en">girder</skos:prefLabel>
</rdf:Description>
```

Figure 23 serves as an example of the alpha version of the final structure, and it shows the elements mentioned above:

- The base URI
- The term URI
- The SKOS properties
- The DC properties
4.4 Linking step

The aim of this part of the work is to generate hyperlinks that connect the glossaries containing the terms extracted from Lynx corpora and the identified existing glossaries from TERMCAT with other knowledge bases and linguistic linked resources that are already part of the Linguistic Linked Open Data cloud.

The reason behind this objective is to easily discover more information about the terms provided by external resources to complement the content of our resource: the relations between terms, usage context, common top concepts, different senses of the same term, translations, etc.

This linking process has also been performed by using OpenRefine: since it was the tool selected for the conversion, the interface and features were already known. The linking service of OpenRefine can be executed in two ways: by means of an SPARQL endpoint or by using an RDF dump of the knowledge base.

For these experiments, only the SPARQL endpoint option has been applied since the three involved knowledge bases offer this kind of access.

These experiments consist in creating links between the five legal glossaries and three relevant knowledge bases: DBpedia, EuroVoc and Babelnet. The aim of these tests is to check the performance of the tool, the relevancy of the terms in these repositories for the legal domain, the needs and future work in this area.

DBpedia\(^{40}\) stores data from Wikipedia by the means of Semantic Web technologies. It is a huge general knowledge base, that provides information about approximately 4 million entities in more than one hundred languages. Many of the entities represented are persons and places, but it seems relevant to test the linking with this knowledge base due to its size. Note: English glossaries have been linked with DBpedia and Spanish glossaries with the Spanish DBpedia.

BabelNet\(^{41}\) is regarded as a multilingual encyclopaedic dictionary but also as a semantic network and a knowledge base. Just like DBpedia, BabelNet also uses data from Wikipedia. However, Wikipedia is only a part of its content: WordNet, Wikidata and Wiktionary are some other resources that contribute to the creation of BabelNet.

Since BabelNet combines general data with lexical information that come from WordNet, it also seems appropriate to analyse how the terms in the glossaries are related to the terms in this net.

EuroVoc\(^{42}\), the multilingual thesaurus of the European Union, contains terminological information in twenty-three languages. The SPARQL endpoint developed by the Semantic Web Company to access the SKOS version of this thesaurus is very convenient to execute the linking experiments. This resource is pertinent since it covers some areas of the legal domain, thus, common terms are expected to be linked.

\(^{40}\) https://wiki.dbpedia.org/
\(^{41}\) https://babelnet.org/
\(^{42}\) http://eurovoc.europa.eu/
4.4 Linking step

4.4.1 Linking Results

Table 10 shows the results of the linking experiments: number of terminological entries linked to each knowledge base and the percentage of the total number of entries contained in each glossary.

<table>
<thead>
<tr>
<th>Glossaries</th>
<th>BabelNet</th>
<th>EuroVoc</th>
<th>DBpedia</th>
<th>Total terms</th>
</tr>
</thead>
<tbody>
<tr>
<td>Labour Law Glossary ES</td>
<td>47</td>
<td>46%</td>
<td>-</td>
<td>8</td>
</tr>
<tr>
<td>Data Protection Glossary EN</td>
<td>70</td>
<td>71%</td>
<td>13</td>
<td>59</td>
</tr>
<tr>
<td>Industrial Standards EN</td>
<td>37</td>
<td>33.94%</td>
<td>-</td>
<td>70</td>
</tr>
<tr>
<td>Termcat Glossary ES</td>
<td>-</td>
<td>-</td>
<td>97</td>
<td>61</td>
</tr>
<tr>
<td>Termcat Glossary EN</td>
<td>-</td>
<td>-</td>
<td>104</td>
<td>118</td>
</tr>
</tbody>
</table>

Tab. 10: Results of the linking tests.

As exposed in this table, the results are average: not too high but neither too low. By analysing the results, the number of links generated is highly dependent on the glossary domain and on the type of knowledge base: for instance, BabelNet performance in linking terms from the Data Protection domain is quite positive, while in the Industrial Standards field is considerably lower.

On the other hand, DBpedia results are quite revealing as well: the lowest percentages belong to the Labour Law glossaries, and there is a huge difference between these results and the number of links generated for more technical domains.

On the whole, the linking process has ended up being virtually manual. OpenRefine means a great help, but disambiguation still represents a serious drawback for which the tool has not found a solution yet. One example of this appeared when linking the term “erasure”, a common concept in the Data Protection domain. All the links suggested by the tool referred to a British music band, and none of the proposed terms had the sense of “data deletion” (see Figure 24).

Fig. 24: Suggested sens of the term “erasure”.

Also, in certain occasions, the tool identifies an exact match between to resources (they have probability 1), but it does not map them automatically (see Figure 25). This means that a reviewer needs to recheck all the entries not to miss any link.
One feature that is particularly annoying is that the tool changes the written representation of the term once it is matched with an external entry. This means that the two entries need to have exactly the same written representation, which prevents from using properties such as skos:majorMatch, skos:minorMatch or skos:relatedMatch. In the legal domain this fact is not very convenient since terms have many variations depending on the jurisdiction or context to which they apply.

Other performance issues have been spotted, regardless of the knowledge base used; they are related to the nature of the terms, since they belong to the legal domain.

When the terms in the language resource that is to be linked are too general, the tool find difficulties to provide a link (see Figure 26). However, the same happens the other way around: when the term is too specific, scarce links are provided. This shows that the information collected in the knowledge bases is driven by usage. Too general terms that everybody should know are not recorded, but too specific terms that only particular groups of people use do not appear either.

Regarding specific issues related to the knowledge base used, with EuroVoc some
unexpected behaviours are occasionally encountered: sometimes the tool does not recognise the properties in which a resource is represented. For instance, the tool not always identifies the `skos:concept` property, which is used to create the links between entries (see Figure 27).

![Reconcile column "prefLabel"](image)

Fig. 27: SKOS property not identified.

One of the main issues with BabelNet is that the links are not generated in the correct language. For instance, when linking the terms of the Industrial Standards glossary, OpenRefine can only find links to terms in three languages: IS, CY and GA, but not in English. However, when looking for the term in English in BabelNet, it does exist. Additionally, the tool was not always able to find any coincidence for compound terms.

On the other hand, in the last stages of this work, an additional linking experiment was performed: since TERMCAT Spanish glossary and Labour Law glossary belong to the same domain, the two resources have been linked together.
Figure 28 represents the final product of this work as a Linguistic Legal Linked Open Data cloud, in which the links generated in this step are also shown (see legend in Table 11).

**Fig. 28:** First approach of the Linguistic Legal Linked Open Data cloud.

<table>
<thead>
<tr>
<th>Legend</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>LOD</td>
<td>Linked Open Data</td>
</tr>
<tr>
<td>LLOD</td>
<td>Linguistic Linked Open Data</td>
</tr>
<tr>
<td>LLLOD</td>
<td>Linguistic Legal Linked Open Data</td>
</tr>
<tr>
<td>DP</td>
<td>Data Protection</td>
</tr>
<tr>
<td>IS</td>
<td>Industrial Standards</td>
</tr>
<tr>
<td>LL</td>
<td>Labour Law</td>
</tr>
<tr>
<td>TC en</td>
<td>Termcat glossary in English</td>
</tr>
<tr>
<td>TC es</td>
<td>Termcat glossary in Spanish</td>
</tr>
</tbody>
</table>

**Tab. 11:** Legend of the LLLOD graph.
4.5 Data portal

As stated in Section 3.6, where the current needs of legal linked data work have been spotted, the major part of the existing resources in this field are not properly documented. In this project, documentation of any generated data is a priority.

With this aim, all the language resources from the legal domain that have been identified and those that have been generated during the whole project are gathered and documented in the Lynx data portal using CKAN technology, available through the Lynx website (http://lynx-project.eu/).

The resources are described according to a set of metadata descriptors that have been collected in two main blocks: information about the dataset and information about the resource.

This distinction is quite significant: within this context, “dataset” makes reference to the whole asset, while “resource” defines each one of the different formats in which the dataset is published (see Tables 12 and 13 for metadata examples).

So, for instance, the UNESCO Thesaurus dataset can be found as two different resources: as a SPARQL endpoint and as a downloadable file in RDF (see Figure 29).

Fig. 29: Example of UNESCO Thesaurus documented in CKAN.
The search application is intended to apply filters such as language, format and jurisdiction. At this moment, there are 29 datasets documented by the OEG in the CKAN, including those linguistic legal linked datasets generated in this project, and some others added by other partners of Lynx consortium. Since Lynx is a 3-years project, this number will increase and having them documented in CKAN will ease their searching and archiving process.
5 Conclusions and future work

In summary, each stage of the whole process has drawn its own results and conclusions. For this reason, they will be exposed separately, with the aim of presenting this section in a clear and organised manner.

Identification of existing resources

The exhaustive search and identification of available language resources from the legal domain has demonstrated that there is a real need for the creation and publication such kind of linguistic assets in structured and open formats.

From the whole set of legal language resources identified in Section 4.1, 36 linguistic assets, 16 of them are already published as RDF (which represents a 44% of the total amount). Six more available resources (17%) are presented in other formats, such as XLS and PDF. The remaining percentage (39%) corresponds to 14 archived resources that are relevant for the domain but cannot be used, either because of their format (some of them are presented as website applications) or because they are not maintained any longer. Some of the latter are European projects, which means they are very valuable resources in this context, but after the project end, they have not been supported anymore.

Taking into account the importance of multilingualism in the legal work within our current society, the percentage of available open legal language resources should be increased.

For this reason, a stage of the future work will be dedicated to the conversion into RDF of many of the identified resources that are in PDF and other unstructured formats since their content is highly relevant for the purposes of this work.

Creation of new resources

In this step, the evaluation of terminological extraction tools has been very useful to keep up to date with the current scenario of such technologies and to identify the needs that may arise when these tools are applied to a larger scale.

Such needs include more precision in the identification of candidate terms, as it is explained in section 4.2.2. Based on the results of the term extraction tasks performed in the creation stage, a 40% of the candidate terms are not correctly identified. This means that a terminologist needs to double-check the terminlists generated to spot those wrong terms and clean the glossary. Such process can be handled when the scope is small, but when this work is part of a European project, the amount of words to review will be too large to be handled manually.

Regarding language technologies, there is a lack of open terminology extraction tools that provide an accurate identification of terms. Besides the fact that most of them add a lot of noise to the resulting list of terms, there is a need of tools that are able to process big corpora as inputs. This means that, even the payable tools do not provide support for big corpora, and this is quite an
important drawback specially for the legal domain, where a huge amount of documentation is being constantly generated.

It would be very convenient to count with a tool that allows to import large amounts of documents automatically, since in Sketch Engine, the user needs to attach each document one by one. For this project, terms have been extracted from corpora composed by sets of 20 files, but in future stages of the Lynx project, greater corpora will be handled.

A possibility to solve this issue would be to use a term extraction tool that is being developed and tested by the UPM, UPM Term Extractor, a Java application to extract terms and relations from scientific paper that was part of Dr. Inventor European Project \[43\] [53]. The advantage of this software is that it can analyse large amounts of documents automatically, in plain text and PDF format, and it returns a bag of words in CSV. However, this resulting bag of words also accumulates noise, and this is why its performance is still being tested.

Another improvement that would be required as part of this legal data management methodology is the treatment of non editable content. Legal documents are specially published as scanned files, since many of the documentation remains available only in physical format. For now, only PDF files generated from editable documents have been handled. However, decisions and agreements (since they usually need to be signed) are normally published as non editable PDF files, thus, a tool to convert such kind of files into editable ones will be required. Currently, several tools for PDF conversion into a machine readable format are available in the market. Some of them are Solid Converter \[44\] and Abbyy FineReader \[45\] and at some point will be used in Lynx. Consequently, an evaluation and testing of the PDF conversion tools will also be performed.

Finally, it would be necessary to add one new process to the methodology proposed. In this work, as it was already stated, the terms have been manually reviewed. Part of this revision consisted in identifying the part of speech of each term in order to create the term ID and, consequently, the URI of each term (as it is explained in section 4.3.1). This process could be automated by using natural language processing technologies applied to part of speech recognition: identifying if the term is a noun (n), verb (v) or adjective (adj) to add this information to the term ID.

\[43\] http://drinventor.eu/
\[44\] https://www.soliddocuments.com/
\[45\] https://www.abbyy.com/en-eu/finereader/
Conversion into RDF

Moving on to the next stage, the conversion process of the resources into RDF also requires some improvements. In the first place, it is important to decide which kind of information is going to be represented, in order to choose the most accurate model and properties.

Currently, glossaries are monolingual termlists. Some entries contain definitions, broader terms and top concepts and all these properties have been represented with the SKOS model. In the future, these glossaries will also contain translations, and they will need to use more complex models to represent linguistic information such as Ontolex.

The Ontolex vartrans module, briefly described hereunder, seems very appropriate to this end, and it was already applied to related projects such as the Terminoteca RDF, also mentioned in section 3.1.1.

An example of an entry contained in Terminoteca RDF, modeled with vartrans can be found in Annexes section (Figure 30) [24].

This model is intended to represent relations amongst entries in different languages, and it makes a distinction between sense relations and form relations:

- **Sense relations** are semantic relations that include terminological and translation relations.

- **Form relations** represent the lexical form of the entry and include morphological and grammatical relations.

This information is necessary to model translations, and SKOS may not be the best vocabulary to represent it. For this reason, the Ontolex vartrans module is proposed as future work, and it will be evaluated and tested to represent translations that will be included in the current glossaries. Since Lynx is an European project involving members from several countries, representing translations is a key step in the project.

Linking step

Similarly, regarding RDF management and linking tools, it would be necessary a greater level of automation in general: from the installation process to the link generation stage. Specifically, OpenRefine offers a simple interface, it is a powerful application, and it is easy to use. However, the linking step is very time-consuming. In a project like Lynx, where large amounts of data are handled, it is not feasible to manually evaluate every linked term like it was done in this work.

In future stages of Lynx project, the applications analysed in the State of the Art (Section 3.5.2) will be tested and used for the conversion into RDF and linking of new datasets that are to be added to the Linguistic Legal Linked Open Data cloud.

As well, several of the resources identified in Section 4.1 are particularly interesting since they present features related to Lynx, such as language, format and domain.
Consequently, they are **valuable datasets to be linked** with the Linguistic Legal Linked Open Data cloud, expanding its scope.

Therefore, the **most immediate step** to continue with this project is to convert (if necessary) and link the following resources:

- German Labour Law Thesaurus\(^{46}\) (already in RDF).
- UNESCO Thesaurus\(^{47}\) (already in SKOS).
- EUGO Glossary\(^{48}\) (conversion required).
- IMF Glossary\(^{49}\) (conversion required).

\(^{46}\) [http://data.lynx-project.eu/dataset/gllt](http://data.lynx-project.eu/dataset/gllt)

\(^{47}\) [http://data.lynx-project.eu/dataset/unesco-thesaurus](http://data.lynx-project.eu/dataset/unesco-thesaurus)

\(^{48}\) [http://data.lynx-project.eu/dataset/eugo-glossary](http://data.lynx-project.eu/dataset/eugo-glossary)

\(^{49}\) [http://data.lynx-project.eu/dataset/imf](http://data.lynx-project.eu/dataset/imf)
6 ANNEXES

CKAN Metadata

<table>
<thead>
<tr>
<th>XLSX</th>
<th>JSON</th>
</tr>
</thead>
<tbody>
<tr>
<td>Title</td>
<td>“field_name”: “title”</td>
</tr>
<tr>
<td>URI</td>
<td>“field_name”: “name”</td>
</tr>
<tr>
<td>Type in the LKG</td>
<td>“field_name”: “lkg_type”</td>
</tr>
<tr>
<td>Type</td>
<td>“field_name”: “type”</td>
</tr>
<tr>
<td>Domain</td>
<td>“field_name”: “domain”</td>
</tr>
<tr>
<td>Identifiers</td>
<td>“field_name”: “identifier”</td>
</tr>
<tr>
<td>Description</td>
<td>“field_name”: “notes”</td>
</tr>
<tr>
<td>Availability</td>
<td>“field_name”: “availability”</td>
</tr>
<tr>
<td>Languages</td>
<td>“field_name”: “language”</td>
</tr>
<tr>
<td>Creator</td>
<td>“field_name”: “creator”</td>
</tr>
<tr>
<td>Publisher</td>
<td>“field_name”: “publisher”</td>
</tr>
<tr>
<td>License</td>
<td>“field_name”: “licence”</td>
</tr>
<tr>
<td>Other rights</td>
<td>“field_name”: “other_rights”</td>
</tr>
<tr>
<td>Jurisdiction</td>
<td>“field_name”: “jurisdiction”</td>
</tr>
<tr>
<td>Date of this entry</td>
<td>“field_name”: “date”</td>
</tr>
<tr>
<td>Proposed by</td>
<td>“field_name”: “partner”</td>
</tr>
<tr>
<td>Number of entries</td>
<td>“field_name”: “total_number”</td>
</tr>
<tr>
<td>Last update</td>
<td>“field_name”: “last_update”</td>
</tr>
<tr>
<td>Dataset organisation</td>
<td>“field_name”: “owner_org”</td>
</tr>
</tbody>
</table>

Tab. 12: Correspondence of dataset metadata in JSON.

<table>
<thead>
<tr>
<th>XLSX</th>
<th>JSON</th>
</tr>
</thead>
<tbody>
<tr>
<td>Name</td>
<td>“field_name”: “name”</td>
</tr>
<tr>
<td>Description</td>
<td>“field_name”: “distribuciones”</td>
</tr>
<tr>
<td>Data format</td>
<td>“field_name”: “format”</td>
</tr>
<tr>
<td>Data access</td>
<td>“field_name”: “lkg_type”</td>
</tr>
<tr>
<td>Open format</td>
<td>“field_name”: “formatoabierto”</td>
</tr>
<tr>
<td>URI</td>
<td>“field_name”: “url”</td>
</tr>
</tbody>
</table>

Tab. 13: Correspondence of resource metadata in JSON.
<table>
<thead>
<tr>
<th>Tool name</th>
<th>Tool format</th>
<th>Input format</th>
<th>Access</th>
<th>Language filter</th>
<th>Output format</th>
<th>Compound terms</th>
<th>Stop words</th>
<th>Additional services</th>
</tr>
</thead>
<tbody>
<tr>
<td>Translated.net</td>
<td>Online</td>
<td>TXT</td>
<td>Free</td>
<td>EN, IT, FR</td>
<td>HTML</td>
<td>Yes</td>
<td>No</td>
<td>Link terms with Google</td>
</tr>
<tr>
<td>VocabGrabber</td>
<td>Online</td>
<td>TXT</td>
<td>Pay</td>
<td>No filter</td>
<td>HTML</td>
<td>No</td>
<td>Yes</td>
<td>Graphical representation</td>
</tr>
<tr>
<td>TermoStat Web</td>
<td>Online</td>
<td>TXT</td>
<td>Free</td>
<td>EN, ES, FR, IT, PT</td>
<td>Multiformat</td>
<td>Yes</td>
<td>No</td>
<td>-</td>
</tr>
<tr>
<td>Sketch Engine</td>
<td>Online</td>
<td>Multiformat</td>
<td>Pay</td>
<td>Multilingual</td>
<td>Multiformat</td>
<td>Yes</td>
<td>No</td>
<td>Multiservice</td>
</tr>
<tr>
<td>Five Filters</td>
<td>Online</td>
<td>TXT</td>
<td>Free</td>
<td>No filter</td>
<td>Multiformat</td>
<td>Yes</td>
<td>Yes</td>
<td>-</td>
</tr>
<tr>
<td>Termine</td>
<td>Online</td>
<td>TXT</td>
<td>Free</td>
<td>No filter</td>
<td>Multiformat</td>
<td>Yes</td>
<td>Yes</td>
<td>-</td>
</tr>
<tr>
<td>Fopple</td>
<td>Downloadable</td>
<td>TXT/PDF</td>
<td>Free</td>
<td>No filter</td>
<td>HTML, TXT</td>
<td>Yes</td>
<td>No</td>
<td>-</td>
</tr>
<tr>
<td>TBXTools</td>
<td>Downloadable</td>
<td>TXT</td>
<td>Free</td>
<td>EN</td>
<td>TXT</td>
<td>Yes</td>
<td>No</td>
<td>Training required</td>
</tr>
<tr>
<td>TermSuite</td>
<td>Downloadable</td>
<td>TXT</td>
<td>Free</td>
<td>Multilingual</td>
<td>TXT</td>
<td>No</td>
<td>Yes</td>
<td>Word analysis</td>
</tr>
</tbody>
</table>

Tab. 14: Term extraction tools comparison.
Fig. 30: Terminesp entry (included in Terminoteca RDF) modelled with Ontolex.
6.1 Accepted abstract for “Law via the Internet” Conference. October 2018.
Towards a Linked Open Data Cloud of Language Resources in the Legal Domain

Patricia Martín Chozas, Elena Montiel Ponsoda, Víctor Rodríguez-Doncel
Ontology Engineering Group, Universidad Politécnica de Madrid, Spain

Abstract. This paper describes the process of identifying and transforming terminological resources in the legal domain into RDF. A survey of heterogenous legal resources in different languages and applicable to different jurisdictions is made, and a selected group of terminologies is transformed into RDF and published as linked data.

Keywords: language resources, semantic web, linked data, legal knowledge graph

1. Introduction

Most practitioners of the legal profession are pleased to ornate their office with books, and legal dictionaries are never missing in their collections. Legal dictionaries used to be valuable resources in their daily job; but nowadays computers have revoked their usefulness. These computers, nevertheless, still need from language data to work properly, and a new breed of electronic language resources has taken over. In this context, language resources are defined as pieces of structured data in a machine-readable form, comprising corpora, terminologies, thesauri, knowledge bases, lexicons and dictionaries. These resources are necessary to train machine translation tools, to automate software localisation systems or to test the natural language processing algorithms of a speech recognition system, to mention but a few.

Examples of language resources in the legal domain are Jurivoc, a juridical thesaurus for Swiss regulations; the UNESCO thesaurus, which contains terms from various fields including the legal domain; or the STW thesaurus, covering the economy domain. These resources were intended for human consumption, but they have been repurposed to be consumed by machines.

Entries in these databases are naturally connected through hyperlinks within the same resource (a dictionary referring to other entries in the same dictionary), across similar resources (you can jump online from an entry in the Random House dictionary to the equivalent in the Merriam Webster) or even across resources of different nature (a corpus of texts with some of the terms linked to entries in another term database). The value of the resources is much higher when connected. We, humans, like hypertext documents in the Web, which enable us to naturally hop from document to document in a form that is connatural with the way we think. Machines perceive many more advantages, and when data is connected to other pieces of data, it is no longer considered as data but as knowledge. Knowledge graphs are no other thing that a set of connected pieces of information. A knowledge graph is a structure to represent information, where entities are represented as nodes, their attributes node labels and the relationship between entities are represented as edges.

This paper is the first effort towards the construction of a knowledge graph of language resources in the legal domain. The overall objective of the work where this paper is framed is the construction of a Legal Knowledge Graph (LKG) enabling the provision of compliance-
related services. This is the main goal of the H2020 Lynx\(^5\) project, and this contribution is a part of it, specifically focusing on language resources. The rest of the abstract is organised as follows: Section 2 describes the goal of this work, a knowledge graph of language resources in the legal domain, together with a first account of identified assets. Section 3 describes the process of transforming existing language resources and adding them to the graph, together with the future work.

2. **Linked Open Data Cloud of legal language resources**

Many resources in the legal domain can already claim to be connected, like any HTML or XML documents with hyper-references to other documents. However, a good way to describe connected resources on the Web relies on the W3C specifications of the Semantic Web, such as RDF\(^6\), RDFS\(^7\), OWL\(^8\) and SKOS\(^9\). Linked Data \([1]\) is a particularly sound manner of publishing RDF. Linked data is data published according to the Linked Data Principles \([2]\): entities should be identified via unique URIs; the URIs should be HTTP URIs, follow standard web protocols, return useful information about the resource and contain links to other related resources. Publishing data as linked data improves the interoperability of data and enables a new breed of tools for data analysis, comparative law studies of systems, regulation checks, etc.

Datasets published as linked data are part of the *Linked Open Data (LOD) cloud*\(^10\), a diagram representing connected linked data resources. The *Linguistic Linked Open Data (LLOD) cloud*\(^11\) \([3]\) is a subset of the former, restricted to datasets in the linguistic domain. The first objective of the work presented here is the identification of existing linked open data language resources in the legal domain. This *Linguistic Legal Linked Open Data (LLLOD) cloud* shaped here would be the inner core of a broader Legal Knowledge Graph, where other non-RDF documents are also referenced.

Despite the existence of language resource portals, there is no good catalogue focused on language legal resources and identifying the relevant resources in the domain is already a first contribution of this work. In order to identify relevant resources, three different paths were explored: (a) general web search; (b) lookup of resources described in papers from the specialized literature and (c) search in data portals specialized in language resources. The latter includes ELRC-SHARE repository (used for documenting language resources by the European Language Resource Coordination), ReTeLe Catalogue (for language resources in Spain), CLARIN (European research infrastructure for language resources), the OLAC Language Resource Catalogue\(^12\) (unified portal for language resource search) and the ELRA Catalogue (European Language Resources).

Each of the resources of interest was described in terms of (a) a general description; (b) whether the dataset is RDF or not and if it is available as linked data and (c) which other resources were connected to this one. Table 1 shows the initial compilation of resources, whereas Figure 1 illustrates the interconnections between some of them.

---

\(^5\) [http://www.lynx-project.eu/](http://www.lynx-project.eu/)
\(^6\) [https://www.w3.org/RDF/](https://www.w3.org/RDF/)
\(^7\) [https://www.w3.org/TR/rdf-schema/](https://www.w3.org/TR/rdf-schema/)
\(^8\) [https://www.w3.org/TR/owl2-syntax/](https://www.w3.org/TR/owl2-syntax/)
\(^9\) [https://www.w3.org/TR/skos-reference/](https://www.w3.org/TR/skos-reference/)
\(^10\) [http://lod-cloud.net/](http://lod-cloud.net/)
\(^11\) [http://linguistic-lod.org/llocloud](http://linguistic-lod.org/llocloud)
<table>
<thead>
<tr>
<th>ID</th>
<th>Name</th>
<th>Description</th>
<th>Language</th>
</tr>
</thead>
<tbody>
<tr>
<td>iate</td>
<td>IATE EU terminological database.</td>
<td>EU languages</td>
<td></td>
</tr>
<tr>
<td>eurovoc</td>
<td>Eurovoc EU multilingual thesaurus.</td>
<td>EU languages</td>
<td></td>
</tr>
<tr>
<td>eur-lex</td>
<td>EUR-Lex EU legal corpora portal.</td>
<td>EU languages</td>
<td></td>
</tr>
<tr>
<td>conncit-legal-glossary</td>
<td>Connecticut Legal Dictionary Bilingual legal glossary.</td>
<td>en, es languages</td>
<td></td>
</tr>
<tr>
<td>unesco-thesaurus</td>
<td>UNESCO Thesaurus Multilingual multidisciplinary thesaurus.</td>
<td>en, es, fr, ru languages</td>
<td></td>
</tr>
<tr>
<td>library-of-congress</td>
<td>Library of Congress Legal corpora portal.</td>
<td>en</td>
<td></td>
</tr>
<tr>
<td>imf</td>
<td>International Monetary Fund Economic multilingual terminology.</td>
<td>en, de, es languages</td>
<td></td>
</tr>
<tr>
<td>eugo-glossary</td>
<td>EUGO Glossary Business monolingual dictionary.</td>
<td>es</td>
<td></td>
</tr>
<tr>
<td>cdisc-glossary</td>
<td>CDISC Glossary Clinical monolingual glossary.</td>
<td>en</td>
<td></td>
</tr>
<tr>
<td>stw</td>
<td>STW Thesaurus for Economics Economic multilingual thesaurus.</td>
<td>en</td>
<td></td>
</tr>
<tr>
<td>edp</td>
<td>European Data Portal EU datasets.</td>
<td>EU languages</td>
<td></td>
</tr>
<tr>
<td>inspire</td>
<td>INSPIRE Glossary (EU) General terms and definitions in English.</td>
<td>en</td>
<td></td>
</tr>
<tr>
<td>saij</td>
<td>SAIJ Thesaurus Controlled list of legal terms.</td>
<td>es</td>
<td></td>
</tr>
<tr>
<td>calathe</td>
<td>CaLaThe Cadastral vocabulary.</td>
<td>en</td>
<td></td>
</tr>
<tr>
<td>Gemet</td>
<td>GEMET General multilingual thesauri.</td>
<td>en, de, es, it languages</td>
<td></td>
</tr>
<tr>
<td>informea</td>
<td>InforMEA Glossary Monolingual glossary on environmental law.</td>
<td>en</td>
<td></td>
</tr>
<tr>
<td>copyright-termbank</td>
<td>Copyright Termbank Multi-lingual termbank of copyright-related terms.</td>
<td>en, es, fr, pt languages</td>
<td></td>
</tr>
<tr>
<td>gllt</td>
<td>German labour law Thesaurus with labour law terms.</td>
<td>de</td>
<td></td>
</tr>
<tr>
<td>jurivoc</td>
<td>Jurivoc Juridical terms from Switzerland.</td>
<td>de, it, fr languages</td>
<td></td>
</tr>
<tr>
<td>termcat</td>
<td>Termcat Terms from several fields including law.</td>
<td>ca, en, es, de, fr, it languages</td>
<td></td>
</tr>
<tr>
<td>termcoord</td>
<td>Termcoord Glossaries from EU institutions and bodies.</td>
<td>EU languages</td>
<td></td>
</tr>
<tr>
<td>agrovoc</td>
<td>Agrovoc Controlled general vocabulary.</td>
<td>29 languages</td>
<td></td>
</tr>
</tbody>
</table>

Table 1. Some features of relevant language resources for the legal domain.

![Diagram](http://data.lynx-project.eu/)

**Figure 1. Relations between identified resources sorted by format.**

Each of these datasets has been described with the DCAT vocabulary and published in the CKAN-based open data portal of the Lynx project[^13^], where they can be browsed using facets.

(language, domain format, jurisdiction, etc.). Further description of the language resources and the research material of this paper is online\textsuperscript{14}.

3. **Population of the LOD and future work**

Some relevant datasets with language data useful for the legal domain are already RDF and some of them are not (green and grey bubbles in Figure 1). In order to connect disparate RDF resources, and in order to transform non-RDF resources to RDF, the existence of common data models is of great help. Some of the most important data models in this domain are SKOS, Ontolex and NIF. The first represents concepts and ontological relation as is common in thesauri in a structured form. The latter second was conceived to model linguistic information relative to ontologies and expose lexical resources in the Semantic Web. In addition, since Lynx is developed in a multilingual environment, the vartrans module of Ontolex is relevant. NIF is of interest to represent annotations in annotated corpora etc.

In order to transform non-RDF resources to SKOS, Ontolex and NIF, the OpenRefine\textsuperscript{15} tool has been chosen. The SKOS and Ontolex properties have been chosen according to the information that needs to be represented (lexical and terminological information) and to the structure of the RDF resources gathered in order to ease the linking stage. Accordingly, some of the most relevant properties spotted are contained in Table 2.

<table>
<thead>
<tr>
<th>SKOS</th>
<th>Ontolex</th>
</tr>
</thead>
<tbody>
<tr>
<td>skos:prefLabel</td>
<td>ontolex:lexicalEntry</td>
</tr>
<tr>
<td>skos:altLabel</td>
<td>ontolex:canonicalForm</td>
</tr>
<tr>
<td>skos:definition</td>
<td>ontolex:language</td>
</tr>
<tr>
<td>skos:note</td>
<td>ontolex:writtenRep</td>
</tr>
<tr>
<td>skos:broadern</td>
<td>ontolex:sense</td>
</tr>
<tr>
<td>skos:topConcept</td>
<td></td>
</tr>
</tbody>
</table>

Table 2. SKOS and Ontolex properties considered (non-equivalent).

The discovery and growth of the LLLOD cloud is an endeavour hardly started. There are many steps to be taken: the work described here is just a first approach to a higher scope. More datasets need to be identified with their subsequent transformation to RDF. On the other hand, specific language resources for Lynx project are to be generated. Finally, the linking of all the resources needs to be performed, contributing in this manner to the enrichment of the Semantic Web.

**References**


\textsuperscript{14} http://lynx-project.eu/data/language-resources

\textsuperscript{15} http://openrefine.org/
6.2 Accepted abstract for “Encuentros Complutenses en torno a la Traducción” Conference. November 2018.
Choosing the right term when translating is always a hard task, which gets even harder when it comes to legal terms: multiple jurisdictions, different categorizations of legal concepts, huge volume of legislations and changes in law over time are just some hindrances that need to be overcome.

Artificial Intelligence (AI) can be a great help in solving these problems. This discipline covers a wide range of areas, and some of them are quite relevant for language related tasks, such as Natural Language Processing (NLP) or Semantic Web technologies.

NLP involves technology that automate processes related with natural language, such as Automatic Terminology Extraction, that permits to identify the terms in a given set of documents. Once these terms are identified, AI technologies also help in their structuring according to the relations that can be established between and/or among them. This kind of structuring in a sort of conceptual map or knowledge organization system has been a core task in the Semantic Web. Currently, many efforts are devoted to the structuring of terms and its exposure as “open data resources” for its reuse and sharing.

This is the main objective of the so called Web of Data and the Linked Data paradigm, a method to represent, publish and link resources that contain overlapping terms, which has generated a cloud of linked resources of general knowledge and specialized domains as a result.

Usually, Linked Data and Semantic Web technologies are thought to be used by machines, but they are very valuable tools for humans as well, specially for terminologists and translators. An open cloud of linked resources in the legal domain can be of great benefit when trying to understand peculiarities of law, analyse how terms relate to each other, and discover equivalences in other languages.

The creation of a cloud of interlinked legal resources is one of the objectives of the European project Lynx, an H2020 Innovation Action aimed at creating a Legal Knowledge Graph (LKG). This LKG will not only interlink legal documents (legislation, directives, standards, etc.), but also a sound set of multilingual language resources from the legal domain that will be used to annotate and classify the previous mentioned type of data.

Such collection of resources is published as RDF (Resource Description Framework) a format used to easily structure and link information, and it is the first stage of a Linguistic Legal Linked Open Data (LLLOD) cloud. The LLLOD is also expected to be used in machine translation tasks to promote the multilingualism of the graph, but also as a source of documentation for terminologists and translators.

Currently, the LLLOD cloud is just a first approach that links a few existing legal language resources with some new term lists extracted from bilingual legal corpora provided by Lynx. In future stages, the scope will be extended to several countries of the European Union, and more languages and type of data (not only linguistic) will be handled.
References


