Creation and Enrichment of a Terminological Knowledge Graph in the Legal Domain

Patricia Martín-Chozas*[0000−0002−8922−7521]

Ontology Engineering Group, Universidad Politécnica de Madrid**
pmchozas@fi.upm.es
http://oeg-upm.net

Abstract. Domain-specific terminologies are of great use in a number of contexts, such as information retrieval from text documents or supporting humans in translation tasks. However, automated terminology extraction tools usually render plain lists with no additional information (hierarchical relations, definitions or examples of use, amongst others). The output of these tools is very often offered in non-open formats, hampering their reuse and interoperability. Moreover, terminology management tools demand a lot of manual work to curate and enrich the resources and they do not support the representation of terminological relations beyond broader/narrower. The contributions of this Thesis mitigate these problems by automating the creation of rich terminologies from plain text documents, by establishing links to external resources, and by adopting the W3C standards for the Semantic Web. The proposed method comprises six tasks: refinement, disambiguation, enrichment, relation validation, relation extraction and RDF conversion. We have applied this methodology to two different legal corpora, i.e., contracts and collective agreements. The result of this methodology will be a *Terminological Knowledge Graph* that can be exploited by different Natural Language Processing applications.

Keywords: Terminology Management · Linguistic Linked Data· Knowledge Graphs · Semantic Web.

1 Motivation

Language Resources are a remarkably valuable asset in our current multicultural and multilingual society. They are a building block in the majority of the digital media we use in our daily routines: social media, online news, audiovisual content and online shopping, to mention but a few. These activities are possible thanks

* This work has been partially funded by the H2020 project Prêt-à-LLOD, under grant agreement No 825182, the H2020 project Lynx, under grant agreement No 780602 and a Predoctoral grant by the Consejo de Educación, Juventud y Deporte de la Comunidad de Madrid.

to Natural Language Processing tasks such as Machine Translation, Text Annotation, Document Classification or Question Answering, that demand sound language resources to produce optimal results. We can find those resources all over the web, from *dictionaries* of general language to *terminologies* specialised in different domains: industry, medicine, environment, amongst others.

Why, then, the most well-known terminological resources in the legal domain are still published in physical and non machine readable formats?

These *terminological resources* lose enormous value when isolated: physical glossaries, terminologies in PDF and others (some examples are mentioned in Section 2). To help solve these issues, we propose a methodology to automate the generation of interoperable resources and to imptove terminology management processes. We rely on open Semantic Web formats that allow to publish resources as Linked Data [2]. When published according to the Linked Data principles, the resources can be interlinked as machine-readable data in non-proprietary formats, giving birth to Knowledge Graphs [12]. Such graphs are very useful to induce information by diverse applications, since the information can be accessed through any of the nodes in the graph. Some efforts have already been devoted to this task, transforming conventional terminologies into RDF (Section 2).

However, since the legal field has always been a very conservative domain, we can hardly find resources online and it is even more difficult to find them as part of the Semantic Web. Thus, with this methodology we want to fill in the gap of linguistic legal knowledge on the web by producing sound domain-specific language resources and reusing available resources in the *Linguistic Linked Open Data* cloud[1]. Throughout this document, the output of this workflow will be referred as a *Terminological Knowledge Graph* composed of *Linked Terminologies*.

The rest of the paper is structured as follows: Section 2 describes the current State of the Art regarding related tools and resources, Section 3 lists the Research Questions and Expected Contributions, Section 4 explains the proposed Methodology and Section 5 contains the initial Evaluation Plan and Conclusions.

2 State of the Art

In this section we explore current Terminology Management Approaches (2.1), Traditional Terminological Resources (2.2) and Linked Language Resources (2.3).

2.1 Terminology Management Approaches

Originally, terminology extraction has been manually performed by translators. Even with the help of Computer Assisted Translation (CAT) tools the process is not automatic: translators need to select the specific terms to be stored. For instance, the most famous CAT tool, SDL Trados Studio [18], provides a terminology management extension, MultiTerm[2], that allows the easy reuse, sharing

[1] http://linguistic-lod.org/llod-cloud

[2] https://www.sdl.com/es/software-and-services/translation-software/terminology-management/sdl-multiterm/

and update of terminologies. However, it is a proprietary application that uses its own format (MTF.XML), which hinders the reusability of the generated terminologies by other applications. Other tools, such as GesTerm[3], can handle several types of file formats and even offer collaborative options. The main drawback here is the great amount of manual work that the terminology management requires, specially in huge volumes of data.

On the other hand, tools such as SketchEngine [14] and the Tilde Terminology platform [13] can work with large corpora and automatically extract most frequent terms and keywords. Still, the output is a plain list of terms with no additional information, nor lexical neither terminological.

Even tools, such as the PoolParty Semantic Suite [17], that is specially designed handle language resources in Semantic Web formats and allows the creation of hierarchies involves a lot of manual efforts: terms and relations amongst them need to be individually selected by the user.

2.2 Traditional Terminological Resources

One of the most important resources of this kind, at European level, is *IATE*, the terminological database of the European Union, originally built in TBX (TermBase eXchange format). The terms contained belong to several domains and languages, covering the activities of the European Union (agriculture, politics, sociology, medicine, etc.).

At a national level, *Terminesp* is also a great effort developed by the Spanish Association of Terminology[4]. It contains multilingual terms related UNE Spanish Standards that can be searched through an online portal. A more specific resource are the glossaries from the *Terminología Oberta* service developed by the Catalan Terminological Centre (*TERMCAT*)[5], that also cover very different domains, but mainly at a regional level.

We can find other international projects that offer consolidated terminology resources, focused in specific domains such as Medicine (MedTerms[6]), Biology (Biology Dictionary[7]) or Industry (Insights Glossary[8]).

Traditional Terminological Resources in the Legal Domain. As mentioned before, some of the most valuable terminological assets in the legal domain nowadays are still published in physical formats. This is the case of *Black's Law Dictionary*, a monolingual legal dictionary widely used by translators [3].

However, the great part are published in online portals, such as the *Dudario jurídico de la ONU*[9], developed by the Translation department of the United

[3] https://www.termcat.cat/es/gestores-terminologia
[4] http://www.aeter.org/
[5] http://www.termcat.cat/en
[6] https://www.medicinenet.com/medterms-medical-dictionary/article.htm
[7] https://biologydictionary.net/
[8] https://insights.eventscouncil.org/Industry-glossary/
[9] https://onutraduccion.wordpress.com/pref/dudario-juridico/

Nations, that gives information about the correct usage of a term in different contexts. Similarly, the *United Nation Terminology Database (UNTERM)*[10] provides terminology and nomenclatures used in the work of the United Nations in eight different languages.

2.3 Linked Language Resources

A fundamental remark at this point is the distinction between "RDF Resource" and "Linked Resource". An "RDF Resource" can be isolated, but a "Linked Resource" is published in RDF and interconnected with other resources. Thus, here we will analyse Linked Resources as they are the output of this work.

In this regard, the main resources are known as Linguistic Knowledge Bases that collect general vocabulary from several domains. Some of the most important resources of this kind are BabelNet[11], a large multilingual semantic network automatically generated from various resources, such as WordNet[12], a huge lexical database for English words, or ConceptNet[13] a semantic network that represents common sense knowledge to support textual reasoning.

Some of the resources mentioned in Section 2.2 are exposed as online portals but they have also been published as Linked Data:

- IATE was converted into RDF, following the *lemon* model[14] and linked with the European Migration Network glossary [8].
- Terminesp and TERMCAT glossaries were transformed and linked generating the *TerminotecaRDF* platform [5] [4].

We can find other linked terminology efforts from different domains, such as AgroVoc [6], developed by the Food and Agrigulture Organization, or EcoLexicon [1], that collects environmental knowledge.

Linked Terminological Resources in the Legal Domain. With the aim of enriching the legal knowledge gap in the Semantic Web, some experiments have already converted and linked legal language resources. For instance, the linking of *IATE*, *Creative Common licenses*, documents from the *World Intellectual Property Organization (WIPO)* and other relevant resources [16].

Another significant effort is the publication of *EuroVoc* as Linked Data, following the SKOS vocabulary [10]. This thesaurus is maintained by the Publications Office of the European Commission and it contains a great number of terms from the legal domain. It has been linked with resources such as the *UNESCO* and the *GEMET* thesauri, amongst others. EuroVoc is also available through a *SPARQL endpoint*[15], supported by PoolParty[16].

[10] https://unterm.un.org/UNTERM/portal/welcome
[11] https://babelnet.org/
[12] https://en-word.net/
[13] http://conceptnet.io/
[14] https://lemon-model.net/
[15] https://lynx.poolparty.biz/PoolParty/sparql/Eurovoc4.3
[16] https://www.poolparty.biz/

3 Research Questions and Expected Contributions

Based on our motivation and the needs raised from the state of the art we can formulate the following research questions:

1. *How can terminology management processes be enhanced by the use of Semantic Web technologies?*
2. *Is it possible to guarantee the quality and specificity to the legal domain of the resulting terminological knowledge graph?*
3. *Which applications can benefit from terminological resources in Semantic Web formats?*

Consequently, our main expected contribution is summarised as the *Creation and Enrichment of a Terminological Knowledge Graph in the Legal Domain*. Due to the lack of legal terminological resources in the web in general and in the Semantic Web in particular, we have applied this approach to the legal field, but we propose a domain independent methodology that can be applied to other areas of knowledge. It is comprised by the following sub-contributions:

- Refinement of automatically extracted terms.
- Enrichment of such term lists with disambiguated data from external Knowledge Bases.
- Identification of new terminological relations and validation of existing ones.

4 Research Methodology and Approach

This work is the continuation of a Master Thesis, aimed at building a Linguistic Linked Open Data cloud on the Legal domain (see [7]), that served as the foundation of the current work. This work proposed a semi-automatic approach to create Legal Linked Terminologies relying on proprietary software such as SketchEngine[17] and also open-source applications such as OpenRefine[18]. A remarkable contribution of this work was the exhaustive collection of existing legal language resources performed, that can be found here[19]. The huge amount of manual work involved in managing the datasets found with the above mentioned tools was the definite impulse to research on an automatic workflow.

The suggested approach is composed of six subtasks, as illustrated in Figure 1; some of them are ongoing and others are still pending. The base input is a corpus of documents that needs to be processed through a Terminology Extraction step. This task is out of the scope of the contribution, since it is not the goal of our research: there are already very good terminology extraction algorithms with a high performance (such as TTF-IDF or CValue).

[17] https://www.sketchengine.eu/
[18] http://openrefine.org/
[19] http://data.lynx-project.eu/dataset?organization=oeg

We have, however, worked on linguistic patterns to adapt an open source extraction software to the legal terminology [15]. Consequently, the input of our workflow is a raw list of terms previously extracted.

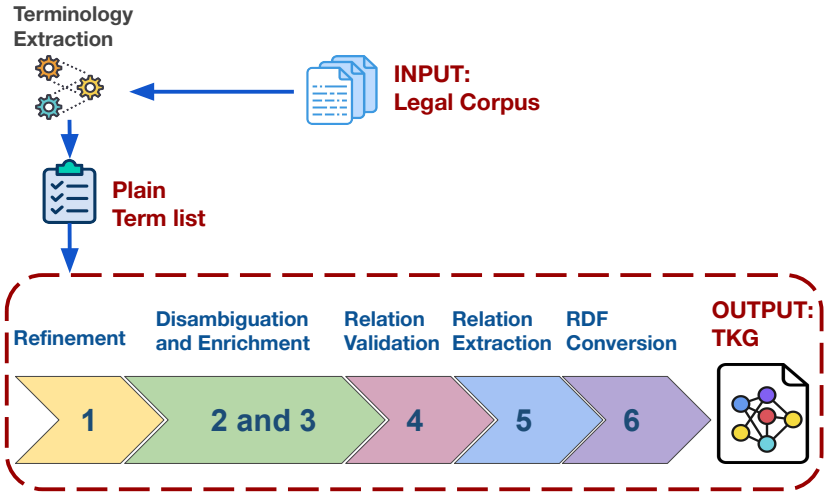

Fig. 1. Terminological Knowledge Graph Generation Workflow

4.1 Task 1: Refinement.

The idea is to implement an open source Automatic Terminology Extraction tool to the workflow, such as TermSuite[20] or JATE [19] so we can directly process corpora. After analysing the raw output of several terminology extraction tools [7], we have noticed that they tend to include noisy terms that need to be filtered. Consequently, we propose a series of automatic refinement suggestions that include lemmatization, removing non terminological structures, removing duplicates, unifying caps, creating top concepts (such as "business" as the top concept of "business partner", "business unit", etc.) and removing Named Entities (such as "Ms Robertson").

4.2 Tasks 2 and 3: Disambiguation and Enrichment.

Once the terms lists are filtered, they can be enriched with additional information by querying external knowledge bases (IATE, Wikidata, EuroVoc, for instance).

However, we first need to make sure that the source term in our terminology and the target term in the queried knowledge base refer to the same lexical sense or concept; this is, terms need to be *disambiguated*. For this task, we are

[20] http://termsuite.github.io/

planning to test existing Word Sense Induction and Disambiguation tools, such as the one developed by Semantic Web Company. We are also researching on disambiguation techniques based on sense embeddings, such as BERT [9].

The idea is to generate sense embeddings from the source and the target terms and compare them: if both vectors are similar, then we assume they refer to the same sense, link both terms and extract relevant information such as translations, synonyms, related terms, etc.

4.3 Task 4: Relation Validation.

In previous enrichment experiments using Wikidata[21] as the external knowledge base, we noticed many issues concerning the data collected under the *also known as* property. The data gathered under this property should be aliases[22] (spelling variants, scientific names and nicknames) and should be categorised as *synonyms* of the source term. However, in many occasions we found broader, narrower and related terms contained under this property, so we have developed a series of axioms to verify each type (Table 1). In this step, we also need to query a second knowledge based specialised in linguistic data, such as ConceptNet, BabelNet or WordNet (see Figure 2).

Table 1. Axioms for inducing semantic relations between alternative labels (A) of a term (T) using term synonyms (S)

Axiom	Induction
$\|T\| = \|A\| \wedge [\forall t_j \in T, \exists! a_i \in A, t_j = a_i \vee a_i \in S_{t_j}]$	T and A are synonyms
$\|T\| < \|A\| \wedge \forall t_j \in T, \exists a_i \in A, t_j = a_i \vee a_i \in S_{t_j}$	A is a narrower term of T
$\|T\| > \|A\| \wedge \forall a_i \in A, \exists t_j \in T, a_i = t_j \vee a_i \in S_{t_j}$	A is a broader term of T
$\exists t_j \in T, a_i \in A, t_j = a_i \vee S_{t_j} \in A$	T and A are related

4.4 Task 5: Relation Extraction.

In this task, our aim is to discover which terminological relations can be found under the `related` property assigned in the previous stage. As an example, in Table 2 we have identified terminological relations for some possible "related" terms of *employment agreement*.

We have already made some research on the state-of-the-art techniques for relation extraction that include lexico-syntactic patterns such as verbal patterns, definitional patterns and knowledge patterns. The use of lexical markers is also a common approach.

[21] https://www.wikidata.org/wiki/Wikidata:Main_Page
[22] https://www.wikidata.org/wiki/Help:Aliases

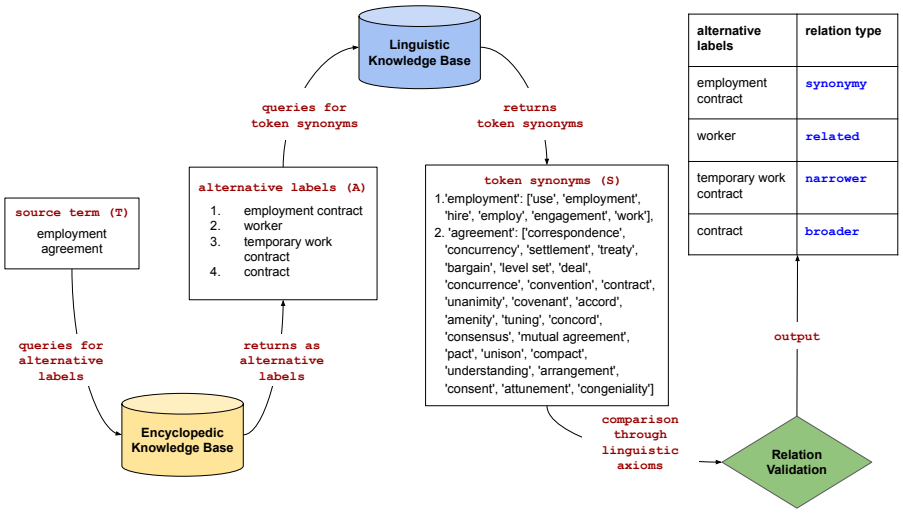

Fig. 2. Relation Validation Example

Table 2. Example of Legal Terminological Relations.

Term 1	Legal Relation	Term 2
Employment Agreement	`signed by`	Worker
Employment Agreement	`negotiated with`	Company
Company	`provides`	Service
Worker	`earns`	Salary
Worker	`works`	Overtime

4.5 Task 6: RDF conversion.

The terminologies are being represented using the SKOS vocabulary[23] since it is an intuitive model whose properties can be used to represent most of the term attributes (`skos:concept`, `skos:prefLabel`, `skos:altLabel`, `skos:description`, `skos:broader`, `skos:narrower` and `skos:related`). However, we still need to research on additional RDF models to represent the properties to be extracted in Task 5 (see Table 2).

The resulting Terminological Knowledge Graph will be serialised as JSON-LD[24], since it is an easy format both for human and machines to interoperate. Our first conversion experiments were done by applying an ad-hoc script; however, to avoid scalability issues, we are researching on mapping language tools that interpret the RDF Mapping Language (RML [11]) and have already been successfully used to transform semi-structured data into Knowledge Graphs.

[23] https://www.w3.org/TR/swbp-skos-core-spec/
[24] https://json-ld.org/

On the other hand, we are considering different ontologies, such as the PROV-O[25] and the Web Annotation Ontology[26] in order to keep track of the provenance of the data.

5 Evaluation plan and Conclusions

5.1 Evaluation

It comprises one of the main challenges of this work, since the most appropriate evaluation should be user based, involving the people for whom the application is intended; in this case, translation, law professionals, students and small enterprises. The issue here is that users need to evaluate the final tool, thus, middle evaluations are more difficult to perform.

In this thesis, we can find two main objects of evaluation: on the one hand, we need to evaluate the Linked Terminology Creation Workflow proposed against other terminology management applications; and on the other hand, we need to evaluate the output, this is, the Terminological Knowledge Graph. For both of them, we will keep track of the data related to the task completeness, efficiency, effectiveness and quality of the result.

The latter can be assessed through a task-based evaluation. This is, we could implement the Terminological Knowledge Graph into a NLP task (such as Question Answering or Machine Translation) and check if its performance is higher than using regular terminological resources (if any).

Additionally, we need to evaluate the maintenance of the Knowledge Graph, this is: research on how to keep the information of the graph updated during the time. An additional task devoted to this aim should be added to the pipeline.

5.2 Conclusions

On the whole, this work remarks 1) the need of great improvements in current terminology management approaches, 2) the usefulness of publishing language resources in Semantic Web formats and 3) specially in the legal domain. Since, even the most used terminology management tools involve a great amount of manual work (as mentioned in Section 2), we propose an automatic workflow to generate linked terminologies directly from corpus. By applying an automatic terminology extraction tool, we work over a plain term lists in 6 different tasks: Refinement, Disambiguation and Enrichment, Relation Validation, Relation Extraction and RDF Conversion. As result, we get linked terminologies that build a Terminological Knowledge Graph.

We want to emphasize in the Relation Extraction task, since we have also spotted a gap on the enrichment of terminologies with the most usual relations amongst terms. Our hypothesis is that this kind of information could improve the results of certain Natural Language Processing tasks.

[25] https://www.w3.org/TR/prov-o/
[26] https://www.w3.org/ns/oa

Bibliography

[1] Araúz, P., A. Reimerink, and P. Faber (2011). Environmental knowledge in ecolexicon. In *Computational Linguistics-Applications Conference*, Number 14, pp. 9.

[2] Berners-Lee, T. (2011). Design issues: Linked data (2006). *URL http://www. w3. org/DesignIssues/LinkedData. html*.

[3] Biel, Ł. (2008). Legal terminology in translation practice: dictionaries, googling or discussion forums. *SKASE Journal of Translation and Interpretation 3*(1), 22–38.

[4] Bosque-Gil, J., J. Gracia, and A. Gómez-Pérez (2016). Linked data in lexicography. *Kernerman Dictionary News 24*, 19–24.

[5] Bosque-Gil, J., E. Montiel-Ponsoda, J. Gracia, and G. Aguado-de Cea. Terminoteca rdf: a gathering point for multilingual terminologies in spain. *TERM BASES AND LINGUISTIC LINKED OPEN DATA*, 136.

[6] Caracciolo, C., A. Stellato, A. Morshed, G. Johannsen, S. Rajbhandari, Y. Jaques, and J. Keizer (2013). The agrovoc linked dataset. *Semantic Web 4*(3), 341–348.

[7] Chozas, P. M., E. M. Ponsoda, and V. Rodríguez-Doncel (2017). Towards a linked open data cloud of language resources in the legal domain.

[8] Cimiano, P., J. P. McCrae, V. Rodríguez-Doncel, T. Gornostay, A. Gómez-Pérez, B. Siemoneit, and A. Lagzdins (2015). Linked terminologies: Applying linked data principles to terminological resources.

[9] Devlin, J., M.-W. Chang, K. Lee, and K. Toutanova (2018). Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*.

[10] Díez, L. A., B. Pérez-León, M. Martínez-González, and D.-J. V. Blanco (2010). Propuesta de representación del tesauro eurovoc en skos para su integración en sistemas de información jurídica. *Scire: representación y organización del conocimiento 16*(2), 47–51.

[11] Dimou, A., M. Vander Sande, P. Colpaert, R. Verborgh, E. Mannens, and R. Van de Walle (2014). Rml: A generic language for integrated rdf mappings of heterogeneous data. *Ldow 1184*.

[12] Ehrlinger, L. and W. Wöß (2016). Towards a definition of knowledge graphs. *SEMANTiCS (Posters, Demos, SuCCESS) 48*.

[13] Gornostay, T. (2010). Terminology management in real use. In *Proceedings of the 5th International Conference Applied Linguistics in Science and Education*, pp. 25–26.

[14] Kilgarriff, A., V. Baisa, J. Bušta, M. Jakubíček, V. Kovář, J. Michelfeit, P. Rychlỳ, and V. Suchomel (2014). The sketch engine: ten years on. *Lexicography 1*(1), 7–36.

[15] Martın-Chozas, P. and P. Calleja (2018). Challenges of terminology extraction from legal spanish corpora.

[16] Rodríguez-Doncel, V., C. Santos, P. Casanovas, and A. Gómez-Pérez (2015). A linked term bank of copyright-related terms. In *JURIX*, pp. 91–100.

[17] Schandl, T. and A. Blumauer (2010). Poolparty: Skos thesaurus management utilizing linked data. In *Extended Semantic Web Conference*, pp. 421–425. Springer.

[18] Walker, A. (2014). *SDL Trados Studio–A Practical Guide*. Packt Publishing Ltd.

[19] Zhang, Z., J. Gao, and F. Ciravegna (2016). Jate 2.0: Java automatic term extraction with apache solr. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16)*, pp. 2262–2269.