Semantic Web 1 (0) 1–5
IOS Press


# TermitUp: Generation and Enrichment of Linked Terminologies

Patricia Martín-Chozas [a,*], Karen Vázquez-Flores [a], Pablo Calleja [a], Elena Montiel-Ponsoda [a] and Víctor Rodríguez-Doncel [a]
[a] *Ontology Engineering Group, Universidad Politécnica de Madrid, Madrid, Spain*
*E-mails: pmchozas@fi.upm.es, kvazquez@delicias.dia.fi.upm.es, pcalleja@fi.upm.es, emontiel@fi.upm.es, vrodriguez@fi.upm.es*

**Editors:** Julia Bosque-Gil, University of Zaragoza, Spain; Milan Dojchinovski, Czech Technical University in Prague, Czech Republic; Philipp Cimiano, Bielefeld University, Germany
**Solicited reviews:** John McCrae, NUI Galway, Ireland; Andon Tchechmedjiev, IMT Mines Alès, France; Four Anonymous Reviewers

**Abstract.** Domain-specific terminologies play a central role in many language technology solutions. Substantial manual effort is still involved in the creation of such resources, and many of them are published in proprietary formats that cannot be easily reused in other applications. Automatic term extraction tools help alleviate this cumbersome task. However, their results are usually in the form of plain lists of terms or as unstructured data with limited linguistic information. Initiatives such as the *Linguistic Linked Open Data cloud (LLOD)* foster the publication of language resources in open structured formats, specifically RDF, and their linking to other resources on the Web of Data. In order to leverage the wealth of linguistic data in the *LLOD* and speed up the creation of linked terminological resources, we propose TermitUp, a service that generates enriched domain specific terminologies directly from corpora, and publishes them in open and structured formats. TermitUp is composed of five modules performing terminology extraction, terminology post-processing, terminology enrichment, term relation validation and RDF publication. As part of the pipeline implemented by this service, existing resources in the *LLOD* are linked with the resulting terminologies, contributing in this way to the population of the *LLOD* cloud. TermitUp has been used in the framework of European projects tackling different fields, such as the legal domain, with promising results. Different alternatives on how to model enriched terminologies are considered and good practices illustrated with examples are proposed.

Keywords: Terminology Generation, Terminology Enrichment, Linguistic Linked Data, Multilingualism

## 1. Introduction

International institutions have become major producers of *multilingual terminology databases*, understood as resources that account for the specialised words used in a particular field in multiple languages. Since its foundation, the European Union has maintained initiatives to cater for the collection, maintenance and creation of terminologies, thesauri or vocabularies, to cover their internal communication needs and to support translators. Some of the best known resources are available from TermCoord[1] (*Terminology Coordination Unit of the European Parliament*), in charge of the interinstitutional terminology database IATE[2] (*InterActive Terminology for Europe*) since 2004, or the EU Vocabularies site[3], maintained by the Publications Office, that is also in charge of the upkeep of the multilingual thesaurus EuroVoc[4].

The creation and curation of such vocabularies has not only supported translators, documentalists and le-

*Corresponding author. E-mail: pmchozas@fi.upm.es.

[1] https://termcoord.eu/
[2] https://iate.europa.eu/
[3] https://op.europa.eu/en/web/eu-vocabularies
[4] http://eurovoc.europa.eu/

gal drafters at EU institutions, but has also become a reference for translators and language professionals outside the EU. Nowadays, curated language resources have proven to be more relevant than ever in light of natural language processing (NLP) tasks that rely on sound linguistic data. For example, query expansion using WordNet[5], the well-known English lexicon [1], disambiguation based on BabelNet[6], a multilingual encyclopedic dictionary [2] and text classification applying DBpedia[7], the semantically structured version of the Wikipedia [3], to mention but a few.

Initiatives such as the *Linguistic Linked Open Data* cloud[8] (henceforward *LLOD*) are focused on collecting and publishing language resources in Semantic Web formats according to the Linked Data principles [4]. When developing NLP services, one of the main challenges is finding language resources on a certain subject area with acceptable quality and ready to be reused, as revealed, for example, in previous experiments on summarisation or machine translation enhanced with terminological resources [5] [6] [7]. Consequently, our motivating scenario is focused on assisting users with different backgrounds and expertise levels to face language and related needs (see Figure 1).

In addition, with the surge in technology solutions for the legal domain, in what is called LegalTech or RegTech, such challenges have become even bigger, since resources of this sort tend to be scarce, private to companies, published in unstructured formats, or no longer available (e.g. the legal multilingual WordNets built in the LOIS project [8], the LexALP term bank on spatial planning and sustainable development [9], or the European legal taxonomy syllabus on consumer protection law [10]). From those resources that have open licenses, such as EuroVoc, most have a wider scope and do not exhaustively cover a specific area of law, or, on the contrary, may only cover a particular sub-area of law (such as the International Labour Organisation Thesaurus[9]); and others are only available in one language or language pair (see abundant examples of terminologies in EuroTermBank[10] project, now eTranslation TermBank and the Wolters Kluwer Thesaurus of Labour Law in German[11]). Therefore, though highly valuable, these resources share some common drawbacks: they usually fall short of covering the specific terminological needs of a certain project or company, are not in the languages of interest, cannot be easily reused or integrated in a new application, and are sometimes only available under request.

With the aim of palliating the need for multilingual terminological resources of a specific domain or project, leveraging resources already available in the *LLOD*, we have devised a method to automatically cover the whole life cycle of the terminology creation process. Our contribution, *TermitUp*, puts together pieces of language technology previously isolated, and improves them to build a pipeline that, taking as input a domain specific corpus in one language, generates as output a multilingual terminology semantically enriched with data from the *LLOD* and published in open formats. The specific subprocesses of the method proposed include terminology extraction, terminology postprocessing, terminology enrichment, relation validation and RDF publication.

Henceforth, the paper is structured as follows: Section 2 presents relevant previous work; Section 3 exposes the linguistic foundations supporting the development of TermitUp; Section 4 lists the application requirements; Section 5 describes every component of TermitUp architecture; Section 6 exposes its current and potential impact; Section 7 contains the discussions that have arisen throughout the development and Section 8 summarises the conclusions and future work.

## 2. Related Work

This section covers previous work related to the different processes mobilised in our system, namely, automatic terminology extraction, modern terminology management tools and semi-automatic terminology enrichment approaches (2.1). We also review existing language resources in RDF and the modelling approaches they follow (2.2).

### 2.1. Terminology-related technology

There is a wide variety of ready-to-use terminology extraction tools, both proprietary (such as SDL MultiTerm Extract[12], TesaurVai[13] and SketchEngine[14]) and

[5] https://en-word.net/
[6] https://babelnet.org/
[7] https://dbpedia.org/
[8] https://linguistic-lod.org/
[9] https://metadata.ilo.org/thesaurus.html
[10] https://www.eurotermbank.com/
[11] https://joinup.ec.europa.eu/solution/wkd-thesaurus-labour-law
[12] https://www.sdl.com/software-and-services/translation-software/terminology-management/sdl-multiterm/
[13] https://www.dail.es/shop/en/
[14] https://www.sketchengine.eu/

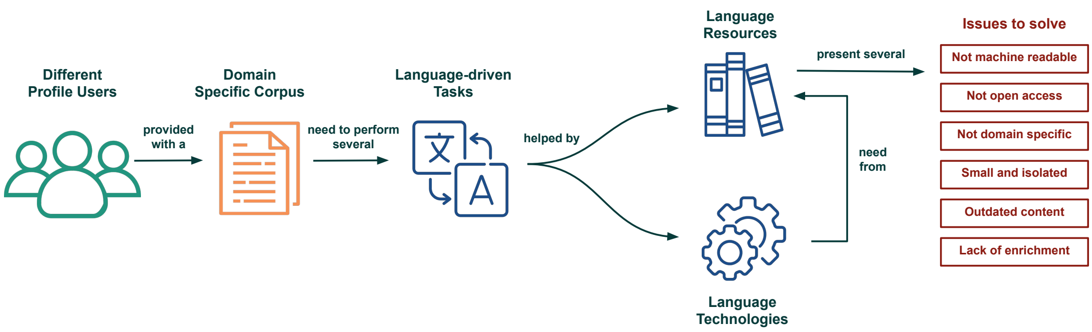


Fig. 1. Motivating scenario for the development of TermitUp.

open source (such as TermSuite[15], TermoStat Web[16] and FiveFilters[17]). There are also implementations of state-of-the-art extraction algorithms, over corpora and over individual documents, such as RAKE [11], JATE [12] or TBXTools [13]. Usually, the main purpose of these tools is to generate plain lists of terms with information about their frequency in the corpus, but no additional linguistic data. Recent approaches are also trying to extract multilingual terminology across domain using transformers, which is a great step forward within the area [14].

More comprehensive terminology management tools integrate monolingual and multilingual term extraction as a starting point feature, and offer additional functionalities to enrich the extracted terms. For example, in Tilde's Terminology platform[18] [15], the extracted terms can be enriched with candidate translations obtained from external resources; SketchEngine[19] [16] identifies collocates for the extracted terms from the source corpus; PoolParty[20] [17] allows the manual creation of hierarchies and the manual linking to resources such as DBpedia[21]; Saffron[22] [18] suggests hierarchical relations between terms, to be afterwards supervised, and VocBench[23] [19] [20] allows the collaborative manual edition of vocabularies.

With regard to semi-automatic terminology enrichment, we also find several approaches in the literature. In [21], the enrichment consists of adding terms to a source thesaurus by exploiting parallel corpora. In [22], WordNet is used to establish *hierarchical* relations between the source terms. Oliveira and Gomes [23] propose a method to automatically enrich a Portuguese thesaurus with synonyms extracted from dictionaries. Some efforts have also been devoted to further specialise the *related to* relation that is common in thesauri with specific semantic relations, as in [24]. In the reviewed works, the scope of the proposed solutions has been limited to one aspect of the terminological resource (synonyms), one external resource (WordNet), or one specific language or language pair. In any case, these efforts deal with one specific feature of the resource or for certain languages, that cannot always be easily extrapolated to other domains or languages.

### 2.2. *Language resources in the Semantic Web*

Concerning existing language resources published in RDF, general domain resources are the most valuable assets in the *LLOD* cloud. WordNet[24], for instance, is a well known general lexicon of the English language that has been converted into RDF following the *lemon* model [25] and linked with many other resources within the cloud. BabelNet is one of the resources that exploits the linked version of WordNet. Combining Wikipedia and other resources, BabelNet constitutes a multilingual semantic network of encyclopedic and language content that covers several domains [26]. The lemon model was also used in the conversion of the Apertium bilingual dictionaries into RDF, a smaller but very relevant work in this area [27].

Apart from the general resources mentioned above, the *LLOD* cloud also gathers some domain specific resources. One of the most important contributions of

[15] http://termsuite.github.io/
[16] http://termostat.ling.umontreal.ca/
[17] https://www.fivefilters.org/term-extraction/
[18] https://term.tilde.com/
[19] https://www.sketchengine.eu/
[20] https://www.poolparty.biz/
[21] https://wiki.dbpedia.org/
[22] https://saffron.insight-centre.org/
[23] http://vocbench.uniroma2.it/
[24] https://wordnet.princeton.edu/

this kind is the RDF dump of IATE, an effort described in [28]. The complete resource is available through a Search API, but not structured in RDF. There have also been efforts to automatically enrich these data [29] with machine translated definitions. IATE offers translations, synonyms and definitions for terms in various domains, but it lacks relations amongst terms.

Some type of term relations are, however, present in *EuroVoc*[25], which gathers data from 21 different domains, with half being closely related to legal activities. Several scientific works are devoted to the conversion of EuroVoc into RDF [30–32] and it is now publicly available through a SPARQL Endpoint hosted by the Publications Office. Although it is not officially part of the *LLOD*, there are several mapping efforts with resources in the cloud. Yet, from the point of view of resources that can be used for NLP tasks, EuroVoc is highly valuable as it contains translations, synonyms and term relations, but lacks other types of linguistic descriptions such as morphosyntactic information or definitions. Also, for domain-specific NLP tasks, frequently, the terms contained are too general, for instance, to process specialised legal documents. Similar issues can be encountered in related resources such as the TheSoz Thesaurus for Social Sciences [33] and the STW Thesaurus for Economics[26], both of them modelled according to SKOS[27]. Unlike EuroVoc, their content is focused on one specific domain, and can be of great help when processing legal documentation. They have, however, an additional limitation: while EuroVoc contains terms in 22 languages, TheSoz is only available for English, French, German and Russian, and STW is bilingual (English-German). The same issue concerns the UNESCO Thesaurus[28], which provides multidisciplinary terminology in English, French, Spanish and Russian. Finally, the International Labour Organisation Thesaurus[29] collects specific terminology for the labour law domain. Unfortunately, terms are only published in English, French and Spanish, synonyms and definitions are scarce, and data is only available under request.

In summary, to ease the creation of terminological resources, we can make use of state-of-the-art terminology extraction tools, although only a few of them provide additional linguistic or semantic data to further describe the terms. To relieve this situation, there have been some approaches pursuing automatic terminology enrichment, yet, they are targeted at one specific type of information, and most of them involve manual efforts. In this paper, we present TermitUp, an automatic approach to generate Multilingual Semantically Enriched Legal Terminologies from corpora in Semantic Web formats. With TermitUp, terms are automatically enriched with translations, term variants or synonyms, definitions, examples of use, information about frequency and hierarchical relations, and are linked with other resources in the *LLOD* cloud.

## 3. Theoretical Underpinnings

The pipeline implemented by TermitUp is in line with the terminographical methods proposed by well-established Terminology theories for the compilation of terminological resources (communicative theory of terminology [34], socioterminology [35], sociocognitive theory of terminology [36] or frame-based theory [37]). In the most common scenario, the starting point in a terminological work is a corpus of specialised texts. The more care taken in constructing the corpus, the better. According to Barrière [38], texts should be domain relevant and contain *knowledge-rich contexts* (a notion defined by Meyer as "sentences that are of interest to terminologists because they contain important terms and *knowledge patterns*", i.e., expressions of semantic relationships [39]). In our approach, the corpus construction task is a manual task assigned to users, who may not be so interested in the knowledge-rich value of texts, but on the relevance of the documents for a certain endeavour.

The next step consists in identifying terminological units in those documents. These can correspond to different part-of-speech (noun, verb, adjective, adverb), and participate in multi-word expressions or phraseological units. Deciding if a lexical unit has a terminological status is not devoid of difficulties. To assist terminologists in this step, several authors propose guidelines in the form of criteria that lexical units have to satisfy to be considered terms [34] [40]. The meaning of a unit is to be discovered in text and constructed through relations to other terminological units. This allows terminologists to derive the conceptual structure underlying those designations, which enables translators or any other language professionals (documentalists, technical writers, subject specialists, etc.) to understand an area of knowledge. Such a structure can

[25]https://publications.europa.eu/en/web/eu-vocabularies
[26]https://zbw.eu/stw
[27]https://www.w3.org/TR/skos-reference/
[28]http://vocabularies.unesco.org/browser/thesaurus
[29]https://metadata.ilo.org/thesaurus/

take the form of an ontology, as suggested in [37], and is the approach taken by the so called *terminological knowledge bases*, as dubbed in [41], in which a knowledge base component is enriched with a linguistic (terminological) component. Some well-known examples of terminological knowledge bases in different areas are GENOMA-KB [42], OncoTerm [43] or EcoLexicon [44].

These theories also propound that terms are to be analysed as used in real communication by experts in the domain, and that this may result in identifying various forms of designations (synonyms or term variants). Variants are to be accounted for in terminological resources, as well as the causes for that variation [45]. Depending on the purpose of the resource at hand, additional linguistic descriptions are also common in terminological resources, namely, source of the term, morphosyntactic information, definition, references to other terms (which can be of different nature, e.g. synonyms, hyponyms, antonyms), usage contexts (that show how the term behaves in real texts), usage notes, or phraseology. Terms are usually assigned to a domain, and all sources from which information has been obtained are referenced, together with other metadata (author, date, reliability degree, etc.).

When considering the multilingual perspective, best practices in terminology work recommend that equivalents in other languages are also collected from domain-specific corpus in the languages of interest, as well as the rest of linguistic descriptions [34]. An exact equivalence relation is assumed when terms in multiple languages are related to a source term, although language and culture differences may be captured in the form of notes. However, previous works on multilingual terminological knowledge bases in the legal domain show how important it is to define culture-specific knowledge as intermediate representations associated with a common shared ontology [46].

Finally, we briefly refer to the theoretical studies (and practical applications) made by terminologists about terminological or conceptual relationships between terms. Conscious of the importance of accounting for such relationships in termbanks, terminographers have characterised them, studied them in particular domains, and created methods for identifying them in corpora. The most important relations in this regard are the so-called hierarchical relationships (hyperonymy-hyponymy and meronymy). However, several non-hierarchical relationships have been intensively studied in some particular domains (cause-effect, entity-function), and others have also been considered for inclusion in terminological resources (antonymy, synonmy, derivative relationships, co-occurrents and collocations). For a nice overview we refer the interested reader to [47].

## 4. Requirements

The development of the first version of TermitUp was guided by a set of requirements derived from the study of existing language technologies, specifically those that deal with terminology, and the observation of their results, as well as from numerous discussions between the linguists, computer scientists and researchers involved in this project.

**R1. Enrichment.** When confronted with domain specific data, there is a need for identifying the specific terms used in texts, as well as their meaning. Plain lists of terms tend not to suffice if they are to be used for annotation, classification or disambiguation and other complex NLP tasks. Definitions, morpho-syntactic information, term variants and explicit relations amongst terms can significantly contribute to improving performance of subsequent text processing tasks.

**R2. Multilingualism.** As already mentioned, international institutions have catered for the creation of multilingual terminologies or thesauri to meet their needs. However, these do not necessarily cover the needs of a company or project in terms of languages, or the purposes of the system being developed. This results in the need for systems that assist in the creation of ad-hoc terminologies for certain language combinations. There have been some initial attempts to developing terminology extractors that work on multilingual corpora, but results are still preliminary.

**R3. Disambiguation.** Although traditional theories to terminology and language planning have backed the approach that the terms in a domain are unambiguous, unique and semantically precise, corpus-based terminology studies have demonstrated that term variation or synonymy is common also in domain specific areas, and that texts may also vary in the degree of specificity. Additionally, external language resources (see Requirement 4) may contain different senses of a term, since they are usually of a general character rather than domain specific. This translates to a necessity for a disambiguation step when matching corpus-extracted terms with external ones.

**R4. Reusability and Standardisation.** Knowledge reuse is the cornerstone of Linked Open Data [4] and the main goal of TermitUp. To meet this objective, this service extracts knowledge from existing resources in the *LLOD* cloud and publishes the resulting terminologies in a structured and open-licensed manner, agreed by the community, so they can be freely reused.

**R5. Data provenance.** When working with texts from a specific domain, it is of utmost importance to guarantee the univocity of the terms managed. Therefore, knowing the source from which each term has been extracted is equally essential, since by knowing these sources, the final user of the terminology has the freedom to choose which term to use depending on the confidence level of such sources. Taking into account that we are dealing with terminologies enriched with heterogeneous external resources, we must maintain traceability not only of the terms themselves, but of each piece of information associated with them: synonyms, translations, definitions, usage examples, etc.

**R6. Open source and easy access.** Following the philosophy of Linked Open Data, we highlight open source as one of the requirements for this service. All the code used will be openly exposed in a Github repository to allow collaboration between users and developers. In addition, TermitUp will be published as a web service for easy integration with other processes.

Throughout this paper, we describe TermitUp functionalities and expose how their specific features comply with each of the requirements above mentioned.

## 5. TermitUp Architecture

With the aim of satisfying the requirements spelled out in the previous section, we present TermitUp, a service to generate domain specific terminologies directly from corpus, enriched with disambiguated terminological data from existing language resources in the *LLOD* cloud. This section describes the five interdependent modules that compose TermitUp architecture.

### *5.1. Module 1: Terminology Extraction*

This module allows to obtain a list of the most representative terms from a given corpus. After analysing and testing several open source automatic term extraction (ATE) tools, and also proprietary software, as mentioned in Section 2, we chose to implement the TBXTools service[30] [48]. TBXTools is a *fast and flexible* tool that offers statistical and linguistic approaches to term extraction. In addition, it is published as a Python library that we could easily implement and modify to satisfy our specific needs (i.e. language and maximum number of tokens per term). The part-of-speech tagging in the linguistic approach is supported by Freeling[31]. However, the performance of the tagger in a preliminary testing phase was not satisfactory compared to other state-of-the-art part-of-speech taggers for Spanish: the application is developed in C++ and its implementation is very time-consuming. Moreover, the results obtained by the statistical method were of good quality, and we decided to rely on the statistical method only.

Originally, TBXTools is intended to process English texts but we fine-tuned the tool to work with Spanish texts (a need arose from our use cases, Requirement 2). We added lists of Spanish stop words and a set of exclusion regular expressions to avoid noisy constructions, which can be consulted in the repository[32].

### *5.2. Module 2: Terminology Post-processing*

Regardless of previously mentioned improvements, we manually reviewed the automatically raw extracted terms and noticed recurrent patterns in Spanish that did not correspond to any multi-word term. For this purpose, we relied on some works that have studied the most common structure of terms in English and Spanish, specifically in the legal domain [34] [49] [50].

Traditionally, nouns were considered the main parts of speech to be included in terminological resources [51], since their main purpose was to label concepts. However, linguistic approaches to terminology argue that terms can belong to different parts of speech (nouns, verbs, adjectives, and sometimes adverbs), often with closely related meanings (for instance, the verb *to contract* and the noun *contract*) [40].

With the objective of filtering common term patters from invalid structures, we designed a post-processing stage in which a *terminology filtering algorithm* relies on part-of-speech annotations to remove structures that do not correspond to valid terms in Spanish. In this regard, a set of 42 linguistic patterns were compiled to detect what we call *non-terminological* structures. Examples of such patterns can be found in Table 1.

Additionally, we also implemented Añotador[33] [52], a service to identify dates and temporal expressions, so

[30] https://sourceforge.net/projects/tbxtools/files/

[31] http://nlp.lsi.upc.edu/freeling/

[32] https://github.com/Pret-a-LLOD/termitup/tree/master/data

[33] https://annotador.oeg.fi.upm.es/

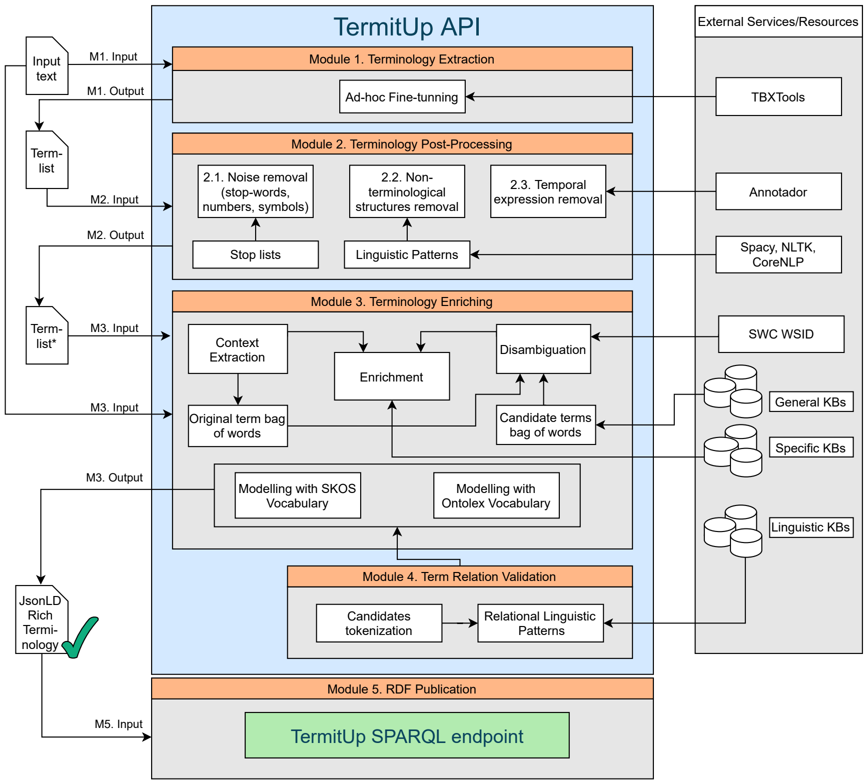


Fig. 2. TermitUp Architecture

that we could also remove them, together with some additional noisy elements.

### 5.3. Module 3: Terminology Enrichment

The next step in this approach is to take full advantage of the information in the *LLOD* relative to the previously filtered terms. Since most of the available resources have a wider scope, either covering several legal areas or general encyclopedic knowledge, a disambiguation process becomes necessary. To this end, we implemented an available word sense disambiguation (WSD) algorithm[34] based on BERT[35].

At this point, we introduce the concept of *sense indicator*, that refers to any word in the surroundings of a term that can be used to disambiguate its meaning.

The algorithm receives as input a *source sense indicator* and several candidate *target sense indicators* from the queried external resources. In TermitUp, the source sense indicator is built by the term $t$ and its surrounding context (up to 100 tokens) from the input corpus $Ct$. For each term we retrieve up to five contexts ($Ct_1 ... Ct_5$). The candidate target sense indicators ($s_1 ... s_n$) consist of any information items related to target terms, such as definitions, synonyms, broader, narrower or related terms, etc.

At first, we assumed that good target sense indicators could be definitions, since definitions contain other relevant words or terms in the domain. For instance: a *training contract* is a particular type of *employment contract* drawn up between an *employer*, a *training organisation* and an *apprentice*. However, we observed that not all the accessed resources contained definitions, so we decided to take every other possible piece of information that could indicate the sense of a term: broader terms, term variants (synonyms) and domain descriptors (see Figure 3). We intentionally avoided using narrower and related terms since often they included terms from neighbouring domains that

[34]https://github.com/semantic-web-company/ptlm_wsid
[35]https://github.com/google-research/bert

Table 1

Examples of Spanish Non-terminological Patterns and Temporal Expressions, and their approximate translation into English for the sake of understanding.

| Exclusion patterns | Examples in Spanish |
|---|---|
| [ADV] | simultáneamente (simultaneously) |
| [ADV] + [ADJ] | inmediatamente posteriores (immediately after) |
| [ADJ] + [ADV] | ininterrumpida inmediatamente (uninterrupted immediately) |
| [NOUN] + [AUX] | partes deberán (parts shall) |
| [NOUN] + [VERB] | consultas corresponderá (enquiries will correspond) |
| [VERB] + [ADJ] | quedar constituida (be established) |
| [VERB] + [NOUN] | produzcan cambios (produce changes) |
| [ADJ] + [ADV] + [ADJ] | objetivas debidamente motivadas (objective duly motivated) |
| [ADJ] + [SCONJ] + [ADV] | negociadora si bien (negotiating as well) |
| [NOUN] + [ADV] + [ADJ] | discriminación tanto directa (discrimination both direct) |
| [NOUN] + [ADV] + [SCONJ] | trabajadores siempre que (workers as long as) |
| [NOUN] + [AUX] + [ADJ] | negociadora estará integrada (negotiating is integrated) |
| [NOUN] + [AUX] + [VERB] | partes deberán negociar (partners should negotiate) |
| [NOUN] + [VERB] + [VERB] | trabajadores podrán acordar (workers could agree) |
| [VERB] + [NOUN] + [ADJ] | concurren causas económicas (economic causes concur) |

| Temporal Expressions in Spanish |
|---|
| 12 de febrero (February 12th) |
| diez semanas (ten weeks) |
| quince días (fifteen days) |
| nueve meses (nine months) |
| febrero de 2012 (February 2012) |
| meses siguientes (following months) |

misled the algorithm. For instance, for the term *promoter*, in the sense of *a person who supports the development of a company*, we get as narrower term *DNA promoter*, as *part of the DNA that starts transcription*.

Table 2 shows an example of the five contexts for the term *hearing* obtained from the input corpus, three sense indicators built with domain descriptors from the queried resource and the resulting *weights*, returned by the WSD implementation. From these weights, the highest refers to the sense that is closest to our domain of interest. From the terms that refer to the sense in question, we can therefore establish a link and enrich our terminology with all the related information available in the queried resources, namely, definitions, translations, synonyms, broader, narrower and related terms. Through this approach, we satisfy Requirement 1: Enrichment; Requirement 2: Multilingualism; and Requirement 3: Disambiguation.

Table 3 lists the *LLOD* language resources exploited and the type of data retrieved from each of them.

### 5.4. Module 4: Term Relation Validation

Some of the resources accessed were originally created and curated by experts. Others, however, were the result of collaborative efforts by users with different levels of expertise. This is why some of the data contained in these resources are not always correct, as it is the case of synonyms and hierarchical relations ob-

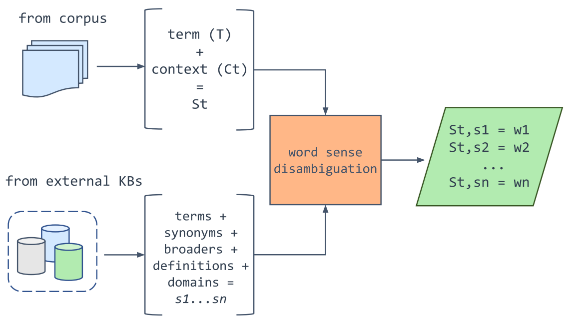


Fig. 3. Representation of the word sense disambiguation workflow

tained, for instance, from Wikidata[36]. The aim of this module is, thus, to check if such relations are correct. Prospective experiments to this module were already published in *LREC* 2020 conference [53], where authors describe the terminology theories that support this work, approach and evaluation of the results.

This approach is inspired by the X-bar theory, that states that the formation of multi-word terms follows a hierarchical structure [34]. The approach suggests a comparison amongst the tokens of terms $t_1$ and $t_2$, and the token synonyms $s1t_1...snt_1$ and $s1t_2...snt_2$ that are retrieved from a linguistic knowledge base. If a synonymy relation is found amongst tokens of two terms, these terms present a terminological relation. The synonyms in this approach were retrieved from

[36]https://www.wikidata.org/

Table 2
WSD example for the term *hearing*, with five different contexts representing the sense of the term, and three candidate sense indicators from the queried knowledge base (IATE in this case). The results show that $s2$ is the closest sense and $Ct4$ the context that better represents it.

| | Context | Results s1 | Results s2 | Results s3 |
|---|---|---|---|---|
| **Ct1** | the difficulty of retaining the hearing date arising from the practical difficulties for the witness | 4.45 | **6.10** | 5.58 |
| **Ct2** | after consideration on the papers by Her Honour Judge Stacey, the ET hearing has since been postponed | 7.44 | **7.46** | 7.02 |
| **Ct3** | it seems that there had been an early case management hearing on 10 April 2017 | 6.22 | **7.79** | 6.88 |
| **Ct4** | the Tribunal may order any person in Great Britain to attend a hearing to give evidence | **7.17** | **7.94** | **7.82** |
| **Ct5** | an application for a witness order may be made at a hearing or by an application in writing to the Tribunal | 6.48 | 7.53 | **7.73** |

| | Senses |
|---|---|
| **s1** | [hearing, parliamentary procedure] |
| **s2** | [hearing, European Union law] |
| **s3** | [hearing, audition, medical science] |

Table 3
List of resources exploited in the legal use case of TermitUp, and the type of information extracted from each of them. All of them are modelled in SKOS and accessed through SPARQL endpoints, except for IATE, whose RDF version is limited and outdated, and its JSON API offers more complete and up-to-date data.

| Resource Name | Type of information available |
|---|---|
| *IATE** | Translations, Synonyms, Definitions, Language Notes and Related Terms |
| *Eurovoc* | Translations, Synonyms, Hierarchical Relations and Related Terms |
| *UNESCO Thesaurus* | Translations, Synonyms, Hierarchical Relations and Related Terms |
| *International Labour Organisation Thesaurus* | Translations, Synonyms, Definitions, Hierarchical Relations and Related Terms |
| *STW Thesaurus* | Translations, Synonyms, Definitions, Hierarchical Relations and Related Terms |
| *Thesoz Thesaurus* | Translations, Synonyms, Definitions, Hierarchical Relations and Related Terms |
| *Wikidata* | Translations, Synonyms, Definitions, Hierarchical Relations and Related Terms |

ConceptNet[37], a large multilingual knowledge graph that brings together data from many open-domain lexical sources (DBpedia, Wiktionary and Open Multilingual WordNet, amongst others). This module can also be used to discover terminological relations amongst the initial term list (see Figure 4).

Additionally, we have implemented a set of rules based on POS-tagging and stemming to generate relations between word forms belonging to the same word family, also known as derivatives. This allows us to group word forms that belong to the same family and gather them under the same concept. Thus, every time we find two terms that follow the patterns *noun-noun, noun-adj, noun-verb, adj-adj, noun-verb* that share the same stem, we generate a *related* relation.

## 5.5. Module 5: RDF Publication

The publication in RDF of the resulting terminological data does not constitute a module of the API itself, but is part of the enrichment module (Module 3), that directly returns a list of files in JSON-LD format for each of the terms processed. Users can choose the vocabulary to represent such files: either SKOS or Ontolex. We consider this choice a fundamental piece of the application, because depending on the future

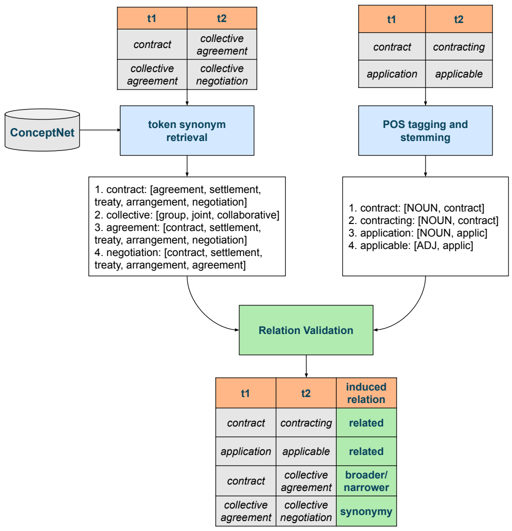


Fig. 4. Relation validation process

[37]http://conceptnet.io/

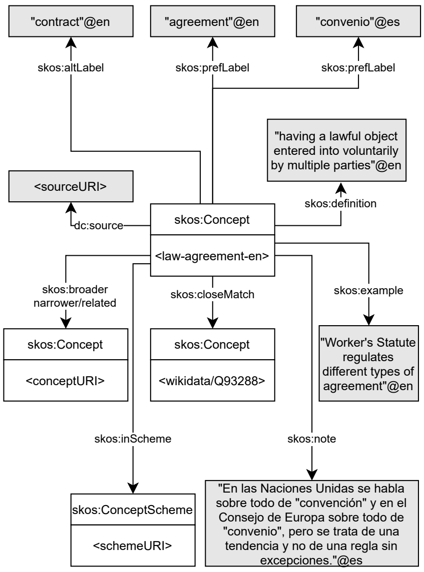


Fig. 5. Example of term modelled with SKOS

application of the terminologies, one model will be more suitable than the other. For example, if the user wants to use this terminology with a tool designed to specifically manage taxonomies, such as PoolParty or VocBench, it is necessary to represent the terminology with SKOS. If, on the contrary, the user intends to enrich the terms with morphological information, then the Ontolex model[38] [54] will be the most appropriate. Figures 5 and 6 exemplify the representation models followed, in which grey boxes represent literals and white boxes represent classes. Some of the white boxes are divided into two parts, where the upper part shows the name of the class and the lower contains some of the properties attached to that class.

Once the user has chosen their preferred RDF vocabulary, the publication module (Module 5) enables the publishing of the results in a Virtuoso Query SPARQL Editor[39] that can be subsequently accessed and queried by the user. The publication is, of course, optional, as the user may want to review the terminology before its publication. The modular architecture of TermitUp allows the human intervention at any point of the pipeline, meaning that the result of each process could be reviewed before starting the next one. In fact, in the future, we would like to developing a terminology editing platform connected to the TermitUp triple store, that allows accessing the terminologies through a user interface, so that users can update them whenever necessary.

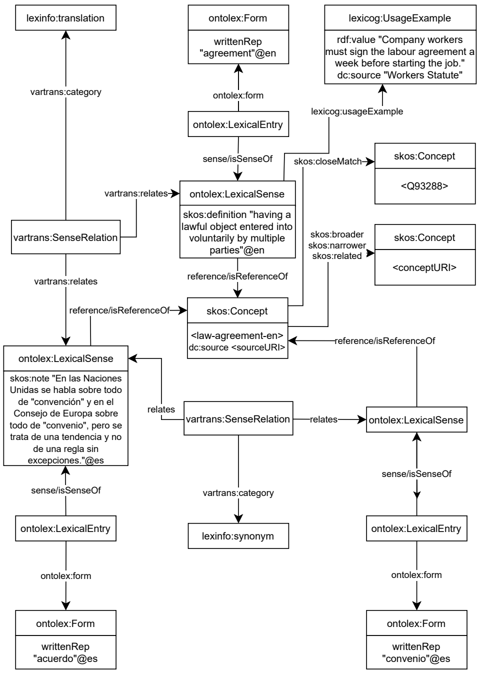


Fig. 6. Example of term modelled with Ontolex

The combination of the exploitation of *LLOD* resources and publication of results in JSON-LD of Module 3, and the publication service represented by Module 5 completely satisfy Requirement 4: Reusability and Standardisation.

## 6. Impact

TermitUp has been developed in the framework of the H2020 Prêt-à-LLOD[40] project, whose objective is to promote the generation and adoption of linguistic technologies that reuse Linked Data. TermitUp contributes to achieving this goal by reducing the human effort necessary to create high quality, rich multilingual terminologies as linked data. In this project, with three pilots of disparate nature, spanning radically different domains such as the pharmaceutical and the e-government ones, TermitUp could be used in a number of different contexts.

[38]https://www.w3.org/2016/05/ontolex/
[39]https://termitup.oeg.fi.upm.es/sparql
[40]https://pret-a-llod.eu/

The main use case of TermitUp has been in the framework of the H2020 Lynx project[41], to produce a labor law terminology, with the intention of improving legal information retrieval tasks –synonyms and hypernyms being of the highest importance. This multilingual terminology (Dutch, English, German and Spanish), after a manual curation made by the domain experts, has been thus validated and published as a SKOS concept scheme. The results are accessible either through the Lynx Terminology platform[42] or downloadable as a static bulk dataset in Zenodo[43]. The main purpose is to contribute to the query expansion process implemented in the Question and Answering Module (SEAR and/or QADocservices), and for navigation purposes amongst documents in different languages. More information about the role of this terminology in the cross-lingual search pilot of Lynx can be found in this deliverable [55].

To evaluate TermitUp's enrichment we have compared this labor law terminology with a gold standard generated from the same corpus (see Table 4). In this gold standard, terms have been manually extracted, semi-automatically enriched and manually reviewed by two Spanish terminology experts. Afterwards, an expert in knowledge management from an international law firm has reviewed and validated the quality of the work. In the context of the project supported by Grupo CPOnet[44], TermitUp is also being used to generate a terminology on crime, where one single punishable event is referred with a surprisingly high number of forms in different language registers.

But the impact of TermitUp goes beyond these domain-specific applications. Its use as a streamlined component in composite workflows suggests a wider range of applications. TermitUp might be used to create user-specific terminologies, contribute to the linguistic analysis of a community, or create more precise vector models, with new features corresponding to the links discovered by TermitUp. In its latest application within the SmarTerp project[45], TermitUp-craft terminologies will support interpreting professionals by providing them extra information on the discourse. The idea of applying TermitUp in this project is based on the analysis of interpreters' needs in terms of domain knowledge presented in [56]. This manual contains a chapter called *Ad hoc Knowledge Acquisition in interpreting*, which explains the documentation phase prior to the interpretation process, highlighting the importance of corpora and terminologies. For this reason, TermitUp fits perfectly as a supporting tool in this documentation phase providing the interpreter with translations, synonyms and related terms to enhance their performance. TermitUp also serves as a means to improve the access and conservation of the glossaries created during the interpretation, helping solve the problem mentioned by Gile: "*Often, because of time pressure, interpreters just write down entries as they encounter them in documents or during the conference, sometimes on sheets of paper they happen to have on hand. They generally do not sort entries manually because of the time this would take. [...] most interpreters either threw away a large proportion of their glossaries prepared for specific conferences or collected them in a disorganized way and lost access to much of the information*". The application of TermitUp in SmarTerp is still preliminary, with a number of challenges related to efficiency pending to be solved, since the project just started.

TermitUp is available in a public GitHub repository[46], as a Python project licensed under Apache License 2.0 terms. The functionality is also available through a HTTP REST API, thus satisfying Requirement 6. These web services are described using OpenAPI [47], and they are running in web servers supported by the Prêt-à-LLOD project. A stable release of the software has also been published in the Zenodo platform[48], favoring the preservation and reproducibility of the research work.

## 7. Discussion

The main limitation found during the development of this service is related to the publication of enriched terminologies in RDF, i.e., to Requirement 5. The objective of this requirement is to maintain the traceability of the data, since the provenance of the information is an essential indicator of its quality. Thus, TermitUp endeavours to store all sources of the collected data.

In the following, we analyse the different type of data collected by the service and the representation possibilities that SKOS and Ontolex offer:

[41] https://lynx-project.eu/
[42] http://lkg.lynx-project.eu/kos
[43] https://zenodo.org/communities/lynx/?page=1&size=20
[44] https://www.grupocponet.com/
[45] https://kunveno.digital/our-proyect/
[46] https://github.com/Pret-a-LLOD/termitup
[47] http://termitup.oeg.fi.upm.es/swagger
[48] https://doi.org/10.5281/zenodo.4461806

Table 4

Comparison of the enrichment numbers of the semi-automatically generated gold standard and the Labor Law terminology automatically generated with TermitUp. We are comparing five types of enrichment and the approximate generation time.

| | **prefLabels** | **altLabels** | **definitions** | **broader/narrower** | **related** | **Estimated Time** |
|---|---|---|---|---|---|---|
| **Gold standard** | 723 | 1229 | 308 | 398 | 162 | ∼ 160 hours |
| **Labor Law Terminology** | 710 | 943 | 264 | 475 | 272 | ∼ 11 hours |
| *Accuracy* | 0.982 | 0.767 | 0.857 | 1.193 | 1.679 | |

– *Terms, synonyms and translations*: In SKOS, they are treated as literals, represented with the properties skos:prefLabel and skos:altLabel, that do not allow to attach any additional information to them. SKOS-XL[49], on the other hand, extends the model to treat these properties as classes, being able to preserve the source. In Ontolex, terms, synonyms and translations are represented as classes, allowing the representation of its source.
– *Context*: the context of a term is treated as an example of how it is used within a text. Therefore, the most suitable property to represent it in SKOS is skos:example (subproperty of skos:note[50]), that allows representing text strings but no additional information. In Ontolex, on the other hand, the Lexicography module [57] considers this need and introduces the lexicog:UsageExample class, that on the contrary, allows representing more information beyond the text itself.
– *Term note*: this is a key element of traditional terminology cards that provides additional information, such as usage recommendations and domain data. Some of the modern language resources do not use term notes anymore, but others still keep them, thus, we consider them valuable pieces of knowledge for language professionals that need to be preserved. In SKOS, term notes can be modeled with skos:note and in Ontolex with ontolex:usage, both object properties pointing to literals. This implies that if we collect term notes from different language resources, we would not be able to model their provenance.
– *Definitions*: the same occurs with definitions, since SKOS vocabulary applies skos:definition, that is also a subproperty of skos:note, therefore an object property that points to a literal. Ontolex does not propose any class for definitions either, and also employs skos:definition. We therefore have the same issue to model its provenance.

Besides the difficulties stated above, we face another modelling decision, since we find different types of sources at different levels. This is, the language resources with which the terms are enriched (i.e. IATE) can be understood as *intermediate sources*, that could be represented with the schema:provider property[51]. Intermediate sources are different from *original sources*, that could be either a corpus (i.e. European Legislation), an organisation (i.e. European Commission), an application (i.e. Definition Extractor) or an individual (i.e. John Doe, terminologist). For their representation, we consider properties such as dc:source and dct:BibliographicResource from DublinCore[52] and the classes prov:Entity, prov:Agent, prov:Person and prov:Organization from PROV ontology[53].

Another discussion that arose from the modelling stage debates was whether the skos:definition (and related documentation properties) should be attached either to the skos:Concept or to the ontolex:LexicalSense. The SKOS specification remains vague in this point, and both approaches are at least syntactically sound – neither skos:definition nor its superproperty skos:note declare a rdfs:domain. This freedom suggests a flexible use which might be suitable to capture some subtleties.

First, when terminological data is transformed from different sources, definitions sometimes seem attached to concepts (e.g. data imported from Wikidata qualifies concepts), sometimes lexical senses (e.g. data imported from WordNet). We suggest the application of skos:definition in a flexible manner, being its subject a skos:Concept or a ontolex:LexicalSense at discretion.

Second, this loosen specification brings about the opportunity to distinguish reference and sense, in *fregean* terms. In his famous essay *Über Sinn und Bedeutung* (1892), Gottlob Frege told apart the reference and the sense of expressions [58]. In this writing, Frege uses the example of Venus: both "the morning star" and "the evening star" refer to the same object, Venus, but

[49]https://www.w3.org/TR/skos-reference/#xl
[50]https://www.w3.org/TR/skos-reference#notes
[51]https://schema.org/provider
[52]https://www.dublincore.org/specifications/dublin-core/dcmi-terms/
[53]https://www.w3.org/TR/prov-o/

the thought they express is rather different. The sense is a mode of presentation, illuminating only a single aspect of the referent. We wonder whether computers can capture these nuances. We can certainly make such an effort, reserving the objective information about Venus for its skos:Concept (e.g. radius = 3000 km), but administer the different subjective perceptions the different components of the synset. Perhaps we want to attribute the ontolex:LexicalSense "Venus" a relatively neutral subjective value related to celestial bodies, and give the ontolex:LexicalSense "morning star" a hotter affective valence, possibly related to a more poetic context. These definitions and affective valences will be necessarily stereotypes, not reflecting subjective values (which are different for each mind), but intersubjective, namely, reflecting common perceptions and images (we refer the reader to [59] for more information about emotional words).

We wonder whether personalized lemonizations will ever be possible, describing the linguistic realities of specific individuals, perhaps inferred from personal big data such as personal email inboxes or alike. But this endeavour is well beyond the scope of this paper; we only stress the opportunity of attributing skos:definition (and other triples) to skos:Concept or ontolex:LexicalSense in the most beneficial manner; in this sense, the *ontoterminology* theory may be a nice point of discussion [60].

We have therefore gathered such ongoing discussions on modelling issues in a proposal for good practices to model terminological resources, published as a *Terminology* draft in the wiki of the Ontology-lexicon Community Group in the W3C[54], where we expose background, motivation and use cases, and suggest complementary elements to the existing models. Such modelling modifications, naturally, need to be agreed by the community before applying them.

## 8. Conclusions and Future Work

The automation in the generation of language resources (specifically, terminological resources) is a challenging task still unresolved. Automatic terminology extraction and terminology management tools provide a good starting point and excellent assistance both for terminology experts and language professionals, but substantial manual effort is still required.

[54]https://www.w3.org/community/ontolex/wiki/Terminology

This contribution intends to lighten such manual efforts, firstly by automating the post-processing step that terminologists usually need to perform over automatically extracted terms, and secondly, by exploiting the wealth of linguistic and terminological knowledge available in the *Linguistic Linked Open Data* cloud. The fact that such resources are published according to Semantic Web standards and open licences contributes to their simple and immediate integration in language technology solutions. However, the majority of them are too general, and do not contain domain-specific terms nor rich linguistic descriptions.

TermitUp helps covering those gaps by extracting and post-processing terms from domain specific corpora, and enriching them with translations, synonyms, definitions, usage notes and terminological relations. Consequently, this application establishes links to the resources exploited, contributing to the population of the *LLOD* with domain expert knowledge. Additionally, the tool offers a module that helps validate the terminological relations retrieved, that sometimes may be imprecise. Finally, the tool structures the resulting enriched terminologies, either following SKOS or Ontolex model; and stores them in a Virtuoso SPARQL Editor so that they can be freely accessed.

If we make a overall comparison with the terminology-related technology presented in Section 2.1, we can notice that TermitUp tackles some issues that they do not observe, which makes TermitUp not a competitor but a complementary application:

- Tilde's Terminology platform extracts terms from corpus and it is able to look for translations in other resources. It, however, does not enrich with definitions, synonyms, usage contexts or relations, and it returns unstructured data.
- SketchEngine is a tool specialised in corpus management. It is also very well known for its terminology extraction algorithm. Although it gives information about term co-occurences and contextual information, the tool does not perform terminology enrichment nor semantic representation.
- PoolParty is a powerful thesaurus management tool that allows creating hierarchical relations amongst terms, representing resources in SKOS and linking them to existing ones such as DBpedia. Still, all the work needs to be manually performed through a user interface. In this case, TermitUp could be used to speed up the terminology generation process and PoolParty would enable the manual revision by experts.

– Saffron was originally a tool for taxonomy extraction. Recent improvements on the tool include terminology extraction, linking to DBpedia and knowledge graph generation. Saffron features are similar to those of TermitUp; it is however intended to work over scientific publications, and the added value is not terminology enrichment as in TermitUp, but "author and content" oriented.
– VocBench is a tool for collaborative management of ontologies and thesauri. It does not generate terminological assets, but helps curate them. As PoolParty, VocBench seems a complementary tool to manage resulting terminologies from the TermitUp workflow.

Furthermore, a remarkable technical benefit of TermitUp is that its development is open source and the community can improve, contribute to or adapt it to their specific uses cases. Also, as it is based on a REST API architecture, TermitUp can be easily integrated with other state-of-the-art technologies or tools.

On the other hand, throughout the development of the service, we have faced several modelling challenges, concretely those related to the provenance of each type of data. With the current vocabularies to model linguistic linked data, not every piece of linguistic information is represented by a class, specifically notes and definitions. This means that no additional information can be added to them, such as the resource from which they have been retrieved. As a consequence, we have discussed and proposed an improvement of the existing models and good practices to accurately structure terminological resources built from heterogeneous data sources to the W3C Ontology-Lexicon Community Group.

During this development, we have also noticed that there is room for improvement in the quality of open (language) knowledge bases available in the *LLOD* - a fact that affects the performance of services relying on them. This is due to the fact that some of the biggest resources, such as Wikidata and ConceptNet, have been semi-automatically built, and their data have not been curated. On the contrary, those manually reviewed, such as KDictionaries' RDF version [61], can only be accessed under permission. We therefore continue pursuing the publishing of high-quality language data in open formats, such as the complete version of IATE RDF, and encourage data owners to do it as well.

Regarding the publishing of the results, an immediate step is to resume the work started in the Terminoteca RDF project [62], whose objective is the creation of a repository of multilingual terminologies. That is, to link different terminologies in a single graph so that they can be queried from a single entry point. Therefore, it seems logical that, since the objective of TermitUp is to generate rich multilingual linked terminologies, the next step would be to publish them in Terminoteca RDF, that would also allow to browse the terms through a graphic interface.

On the other hand, we observed that traditional terminological resources (such as TERMIUM and IATE) do not make explicit the relations that may exist between terms, that are to be inferred by the user from the information contained in definitions or usage notes. Terminological knowledge bases or thesauri, which follow a more conceptual approach, intend to classify concepts in a conceptual structure and include hierarchical relations (broader-narrower term relations), as well as an unidentified type of relation that flags that two terms are somehow related (see "related to" in EuroVoc or Agrovoc). Frame-semantics and other Lexicon driven approaches to terminology (see Section 3) agree on the interest of capturing terminological relations, including *domain-specific relations*, that describe how two terms interact with each other in a given area of expertise. The most generic relations include cause-effect and object-function, for instance.

Consequently, the next version of TermitUp is thought to contain an additional module that allows performing automatic domain-specific relation extraction amongst the terms in the terminology, based on the study of their behaviour in the corpus.

Finally, challenging the current domain-specific application of the tool, we have already two potential projects of very different domains, in which TermitUp will take part: 1) Authors have recently worked jointly with the DFKI research center[55], on the conversion of the terminological base from the Deutsche Bahn (main railway German Company)[56] into Semantic Web formats. This resource lacks Spanish terminological data, and TermitUp will be used to enrich it with Spanish data on the domain. 2) Authors are also involved in a project to transform the main database of Spanish emotional words, Emofinder [63], into a knowledge graph for better access, update and conservation. In this context, TermitUp will not handle *terms* but *words* from the general domain, and it will enrich the resource mainly with translations. In the first case, TermitUp

[55]https://www.dfki.de/
[56]https://www.bahn.de/

will query terminological resources from the transport domain, while in the latter, it will access general purpose resources, adding some important ones such as DBpedia and BabelNet.

## Acknowledgements

This work has been supported by the European Union's Horizon 2020 research and innovation programme through the Prêt-à-LLOD[57] project, with grant agreement No. 825182.

## References

[1] D. Pal, M. Mitra and K. Datta, Improving query expansion using WordNet, *Journal of the Association for Information Science and Technology* **65**(12) (2014), 2469–2478, https://doi.org/10.1002/asi.23143.

[2] R. Navigli and S.P. Ponzetto, Multilingual WSD with Just a Few Lines of Code: the BabelNet API, in: *Proceedings of the ACL 2012 System Demonstrations*, Association for Computational Linguistics, Jeju Island, Korea, 2012, pp. 67–72. https://aclanthology.org/P12-3012.

[3] J. Flisar and V. Podgorelec, Improving short text classification using information from DBpedia ontology, *Fundamenta Informaticae* **172**(3) (2020), 261–297, https://doi.org/10.3233/FI-2020-1905.

[4] C. Bizer, T. Heath and T. Berners-Lee, Linked data: The story so far, in: *Semantic services, interoperability and web applications: emerging concepts*, IGI global, 2011, pp. 205–227, https://doi.org/10.4018/jswis.2009081901.

[5] J. Vivaldi, I. Da Cunha, J.-M. Torres-Moreno and P. Velázquez-Morales, Automatic Summarization Using Terminological and Semantic Resources., in: *Proceedings of the Eleventh International Conference on Language Resources and Evaluation, LREC*, 2010. ISBN 2-9517408-6-7.

[6] F. Schmedding, P. Klügl, D. Baehrens, C. Simon, K. Simon and K. Tomanek, EuroVoc-Based Summarization of European Case Law, in: *AI Approaches to the Complexity of Legal Systems*, Springer, 2015, pp. 205–219, https://doi.org/10.1007/978-3-030-00178-0_13.

[7] M. Arcan, M. Turchi, S. Tonelli and P. Buitelaar, Leveraging bilingual terminology to improve machine translation in a CAT environment, *Natural Language Engineering* **23**(5) (2017), 763–788, https://doi.org/10.1017/S1351324917000195.

[8] D. Tiscornia, The Lois Project: Lexical Ontologies for Legal Information Sharing, in: *Proceedings of of the V Legislative XML Workshop, European Press Academic Publishing, 2007*, pp. 189–204. ISBN 9788883980466.

[9] V. Lyding, E. Chiocchetti, G. Sérasset and F. Brunet-Manquat, The LexALP information system: Term bank and corpus for multilingual legal terminology consolidated, in: *Proceedings of the workshop on multilingual language resources and interoperability*, Association for Computational Linguistics, 2006, pp. 25–31. ISBN 9781932432824.

[10] G. Ajani, G. Boella, L. Di Caro, L. Robaldo, L. Humphreys, S. Praduroux, P. Rossi and A. Violato, The European legal taxonomy syllabus: a multi-lingual, multi-level ontology framework to untangle the web of European legal terminology, *Applied Ontology* **11**(4) (2016), 325–375, https://doi.org/10.3233/AO-170174.

[11] S. Rose, D. Engel, N. Cramer and W. Cowley, Automatic keyword extraction from individual documents, *Text mining: applications and theory* **1** (2010), 1–20, https://doi.org/10.1002/9780470689646.ch1.

[12] Z. Zhang, J. Gao and F. Ciravegna, JATE2.0: Java Automatic Term Extraction with Apache Solr, in: *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC 2016)*, European Language Resources Association (ELRA), 2016. ISBN 978-2-9517408-9-1.

[13] M. Vàzquez and A. Oliver, Improving term candidates selection using terminological tokens, *Terminology. International Journal of Theoretical and Applied Issues in Specialized Communication* **24**(1) (2018), 122–147, https://doi.org/10.1075/term.00016.vaz.

[14] C. Lang, L. Wachowiak, B. Heinisch and D. Gromann, Transforming Term Extraction: Transformer-Based Approaches to Multilingual Term Extraction Across Domains, in: *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, 2021, pp. 3607–3620, https://doi.org/10.18653/v1%2F2021.findings-acl.316.

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

[16] A. Kilgarriff, V. Baisa, J. Bušta, M. Jakubíček, V. Kovář, J. Michelfeit, P. Rychlỳ and V. Suchomel, The Sketch Engine: ten years on, *Lexicography* **1**(1) (2014), 7–36, https://doi.org/10.1007/s40607-014-0009-9.

[17] T. Schandl and A. Blumauer, PoolParty: SKOS thesaurus management utilizing linked data, in: *Extended Semantic Web Conference*, Springer, 2010, pp. 421–425, http://dx.doi.org/10.1007/978-3-642-13489-0_36.

[18] G. Bordea, P. Buitelaar and T. Polajnar, Domain-independent term extraction through domain modelling, in: *The 10th international conference on terminology and artificial intelligence (TIA 2013), Paris, France*, 10th International Conference on Terminology and Artificial Intelligence, 2013. ISBN 978-2-9174-9025-9.

[19] A. Stellato, S. Rajbhandari, A. Turbati, M. Fiorelli, C. Caracciolo, T. Lorenzetti, J. Keizer and M.T. Pazienza, VocBench: a web application for collaborative development of multilingual thesauri, in: *European semantic web conference*, Springer, 2015, pp. 38–53, https://doi.org/10.1007/978-3-319-18818-8_3.

[20] A. Stellato, M. Fiorelli, A. Turbati, T. Lorenzetti, W. Van Gemert, D. Dechandon, C. Laaboudi-Spoiden, A. Gerencsér, A. Waniart, E. Costetchi et al., VocBench 3: A collaborative Semantic Web editor for ontologies, thesauri and lexicons, *Semantic Web* **11**(5) (2020), 855–881, https://doi.org/10.3233/sw-200370.

[21] H. Déjean, E. Gaussier, J.-M. Renders and F. Sadat, Automatic processing of multilingual medical terminology: applications to thesaurus enrichment and cross-language informa-

[57] https://pret-a-llod.eu/

tion retrieval, *Artificial Intelligence in Medicine* **33**(2) (2005), 111–124, https://doi.org/10.1016/j.artmed.2004.07.015.
[22] L. Hollink, V. Malaisé and G. Schreiber, Thesaurus enrichment for query expansion in audiovisual archives, *Multimedia Tools and Applications* **49**(1) (2010), 235–257, https://doi.org/10.1007/s11042-009-0400-y.
[23] H.G. Oliveira and P. Gomes, Towards the automatic enrichment of a thesaurus with information in dictionaries, *Expert Systems* **30**(4) (2013), 320–332, https://doi.org/10.1111/exsy.12029.
[24] Y. Wu, Enriching a thesaurus as a better question-answering tool and information retrieval aid, *Journal of Information Science* **44**(4) (2018), 512–525, http://dx.doi.org/10.1177/0165551517706219.
[25] J. McCrae, C. Fellbaum and P. Cimiano, Publishing and Linking WordNet using lemon and RDF, in: *Proceedings of the 3rd Workshop on Linked Data in Linguistics*, 2014.
[26] R. Navigli and S.P. Ponzetto, BabelNet: Building a very large multilingual semantic network, in: *Proceedings of the 48th annual meeting of the association for computational linguistics*, 2010, pp. 216–225. ISBN 978-1-932432-66-4.
[27] J. Gracia, M. Villegas, A. Gomez-Perez and N. Bel, The apertium bilingual dictionaries on the web of data, *Semantic Web* **9**(2) (2018), 231–240, https://doi.org/10.3233/SW-170258.
[28] P. Cimiano, J.P. McCrae, V. Rodríguez-Doncel, T. Gornostay, A. Gómez-Pérez, B. Siemoneit and A. Lagzdins, Linked terminologies: applying linked data principles to terminological resources, in: *Proceedings of the eLex 2015 Conference*, 2015, pp. 504–517, ISBN 978-961-93594-3-3.
[29] M. Arcan, E. Montiel-Ponsoda, J.P. McCrae and P. Buitelaar, Automatic Enrichment of Terminological Resources: the IATE RDF Example, in: *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC-2018)*, 2018, https://doi.org/10.5281/zenodo.2599942.
[30] L.P. Paredes, J. Rodrıguez and E.R. Azcona, Promoting government controlled vocabularies for the Semantic Web: the EUROVOC thesaurus and the CPV product classification system, *Semantic Interoperability in the European Digital Library* (2008), 111.
[31] M.M. Martínez González, B. Pérez León, M.L. Alvite Díez et al., SKOS en la integración de conocimiento en los sistemas de información jurídica, *Actas del Taller de Trabajo Zoco'09/JISBD* **3**(6) (2009), 56.
[32] L.A. Díez, B. Pérez-León, M. Martínez-González and D.-J.V. Blanco, Propuesta de representación del tesauro Eurovoc en SKOS para su integración en sistemas de información jurídica, *Scire: representación y organización del conocimiento* (2010), https://doi.org/10.54886/scire.v16i2.4015.
[33] B. Zapilko, J. Schaible, P. Mayr and B. Mathiak, TheSoz: A SKOS representation of the thesaurus for the social sciences, *Semantic Web* **4**(3) (2013), 257–263, https://doi.org/10.3233/SW-2012-0081.
[34] M.T. Cabré and M.T.C. i Castellví, *La terminología: teoría, metodología, aplicaciones*, Editorial Antártida/Empúries, 1993.
[35] F. Gaudin, *Socioterminologie*, Publication Univ Rouen Havre, 1993, https://doi.org/10.4000/praxematique.2188.
[36] R. Temmerman, *Towards new ways of terminology description: The sociocognitive-approach*, Vol. 3, John Benjamins Publishing, 2000, http://dx.doi.org/10.4314/lex.v14i1.51442.
[37] P. Faber, Frames as a framework for terminology, *Handbook of terminology* **1**(14) (2015), 14–33, http://dx.doi.org/10.1075/hot.1.02fra1.
[38] C. Barriere and A. Agbago, TerminoWeb: a software environment for term study in rich contexts, in: *International Conference on Terminology, Standardisation and Technology Transfer (TSTT 2006)*, 2006.
[39] I. Meyer, Extracting knowledge-rich contexts for terminography, *Recent Advances in Computational Terminology* **2** (2001), 279, https://doi.org/10.1075/nlp.2.15mey.
[40] M.-C. L'Homme, *Lexical semantics for terminology: an introduction*, Vol. 20, John Benjamins Publishing Company, 2020, https://doi.org/10.1075/tlrp.20.
[41] I. Meyer, D. Skuce, L. Bowker and K. Eck, Towards a new generation of terminological resources: an experiment in building a terminological knowledge base, in: *COLING 1992 Volume 3: The 15th International Conference on Computational Linguistics*, 1992, https://doi.org/10.3115/992383.992410.
[42] M.T. Cabré, C. Bach, R. Estopà, J. Feliu, G. Martínez and J. Vivaldi, The GENOMA-KB Project: Towards the Integration of Concepts, Terms, Textual Corpora and Entities., in: *Proceedings of the Eleventh International Conference on Language Resources and Evaluation, LREC*, 2004. ISBN 2-9517408-1-6.
[43] P. Faber et al., ONCOTERM: sistema bilingüe de información y recursos oncológicos, *La traducción científico-técnica y la terminología en la sociedad de la información* (2002), 177, http://dx.doi.org/10.6035/EstudisTraduccio.2002.10.
[44] P. Faber, P. León-Araúz and A. Reimerink, Representing environmental knowledge in EcoLexicon, in: *Languages for Specific Purposes in the Digital Era*, Springer, 2014, pp. 267–301, http://dx.doi.org/10.1007/978-3-319-02222-2_13.
[45] M.T. Cabré, J. Freixa, M. Lorente and C. Tebé, La terminología hoy: replanteamiento o diversificación, *Organon* **12**(26) (1998). ISBN 9788475964058.
[46] K. Kerremans, R. Temmerman and J. Tummers, Representing multilingual and culture-specific knowledge in a VAT regulatory ontology: Support from the termontography method, in: *OTM Confederated International Conferences" On the Move to Meaningful Internet Systems"*, Springer, 2003, https://doi.org/10.1007/978-3-540-39962-9_68.
[47] M.-C. L'Homme and E. Marshman, Terminological Relationships and Corpus–based Methods for Discovering them: An Assessment for Terminographers, *L. Bowker (Éd.), Lexicography, Terminology, and Translation. Text-based studies in honour of Ingrid Meyer, University of Ottawa Press, Ottawa* (2006), 67–80, http://dx.doi.org/10.2307/j.ctt1ckpgs3.8.
[48] A. Oliver and M. Vàzquez, TBXTools: a free, fast and flexible tool for automatic terminology extraction, in: *Proceedings of the international conference recent advances in natural language processing*, 2015, pp. 473–479. ISSN 1313-8502.
[49] E. Alcaraz and B. Hughes, El español jurídico, *Barcelona: Ariel* (2002). ISBN 978-84-344-1872-1.
[50] E. Alcaraz and B. Hughes, *Legal translation explained*, Routledge, 2014, https://doi.org/10.4324/9781315760346.
[51] M.-C. L'Homme, What can verbs and adjectives tell us about terms?, *Terminology and Knowledge Engineering, TKE 2002. Proceedings* (2002), 28–30, http://dx.doi.org/10.13140/2.1.4075.3927.
[52] M. Navas-Loro and V. Rodríguez-Doncel, Annotador: a temporal tagger for Spanish, *J. Intell. Fuzzy Syst.* **39** (2020), 1979–1991, https://doi.org/10.3233/JIFS-179865.

[53] P. Martín-Chozas, S. Ahmadi and E. Montiel-Ponsoda, Defying Wikidata: Validation of terminological relations in the web of data, in: *The 12th International Conference on Language Resources and Evaluation (LREC)*, 2020. ISBN 979-10-95546-34-4.

[54] J.P. McCrae, J. Bosque-Gil, J. Gracia, P. Buitelaar and P. Cimiano, The Ontolex-Lemon model: development and applications, in: *Proceedings of eLex 2017 conference*, 2017, pp. 19–21, ISSN 2533-5626.

[55] P. Boil, E. Gómez and P. Calleja, Lynx D5.7 Demonstrator for pilot 3, Zenodo, 2020. doi:10.5281/zenodo.4300691.

[56] D. Gile, *Basic concepts and models for interpreter and translator training*, Vol. 8, John Benjamins Publishing, 2009. ISBN 978-9027224323.

[57] J. Bosque-Gil, J. Gracia and E. Montiel-Ponsoda, Towards a Module for Lexicography in OntoLex., in: *LDK Workshops*, 2017, pp. 74–84.

[58] G. Frege, Über sinn und bedeutung, *Zeitschrift für Philosophie und philosophische Kritik* **100** (1892), 25–50. ISBN 9783150195826.

[59] M.M. Bradley and P.J. Lang, Affective norms for English words (ANEW): Instruction manual and affective ratings, Technical Report, The center for research in psychophysiology, University of Florida, 1999.

[60] C. Roche, M. Calberg-Challot, L. Damas and P. Rouard, Ontoterminology: A new paradigm for terminology, in: *International Conference on Knowledge Engineering and Ontology Development*, 2009, pp. 321–326, https://doi.org/10.5220/0002330803210326.

[61] J. Bosque-Gil, J. Gracia, E. Montiel-Ponsoda and G. Aguado-de-Cea, Modelling multilingual lexicographic resources for the Web of Data: The K Dictionaries case, in: *GLOBALEX 2016 Lexicographic Resources for Human Language Technology Workshop Programme*, 2016, p. 65.

[62] J. Bosque-Gil, E. Montiel-Ponsoda, J. Gracia and G. Aguado-de Cea, Terminoteca RDF: a gathering point for multilingual terminologies in Spain, in: *Proceedings of TKE 2016 the 12th International conference on Terminology and Knowledge Engineering*, 2016, pp. 136–146. ISBN 9788799917907.

[63] I. Fraga, M. Guasch, J. Haro, I. Padrón and P. Ferré, EmoFinder: The meeting point for Spanish emotional words, *Behavior Research Methods* **50**(1) (2018), 84–93, https://doi.org/10.3758/s13428-017-1006-3.