

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: First Editor, University or Company name, Country; Second Editor, University or Company name, Country

Solicited reviews: First Solicited Reviewer, University or Company name, Country; Second Solicited Reviewer, University or Company name, Country

Open reviews: First Open Reviewer, University or Company name, Country; Second Open Reviewer, University or Company name, Country

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 –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 TermCoord¹ (*Terminology Coordination Unit of the European Parliament*), in charge of the interinstitutional terminology database IATE² (*InterActive Terminology for Europe*) since 2004, or the EU Vocabularies site³, maintained by the Publications Office, that is also in charge of the upkeep of the multilingual thesaurus EuroVoc⁴.

¹<https://termcoord.eu/>

²<https://iate.europa.eu/>

³<https://op.europa.eu/en/web/eu-vocabularies>

⁴<http://eurovoc.europa.eu/>

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

The creation and curation of such vocabularies has not only supported translators, documentalists and legal 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 on the light of Natural Language Processing (NLP) tasks that rely on sound linguistic data. For example, Query Expansion using WordNet⁵, the well-known English lexicon [1], Disambiguation based on BabelNet⁶, a multilingual encyclopedic dictionary [2] and Text Classification applying DBpedia⁷, 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 to find 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 face language and domain 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 Thesaurus⁹); and others are only available in one language or language pair (see abundant examples of terminologies in EuroTermBank¹⁰ project, now eTranslation TermBank and the Wolters Kluwer The-

saurus of Labour Law in German¹¹). 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 each component of *TermitUp* architecture; section 6 exposes the its current and potential impact; section 7 contains the discussions aroused throughout the development and section 8 summarises the conclusions and future work.

2. Related Work

This section attempts to cover previous work related to the different processes covered by the 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 Multi-

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>

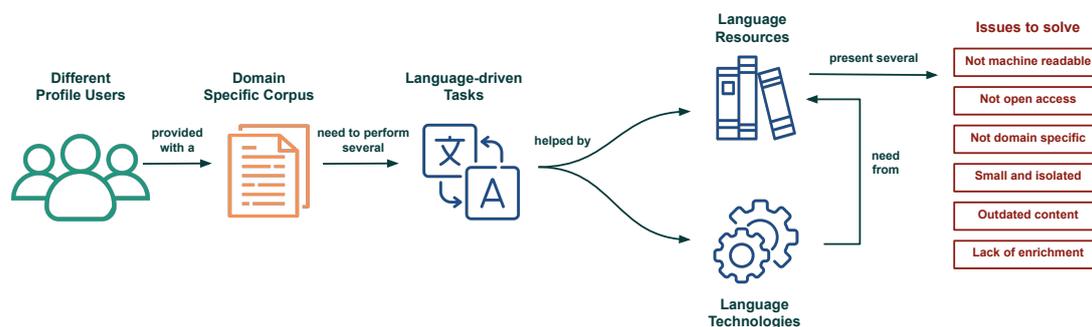


Fig. 1. Diagram representing the motivating scenario for the development of TermitUp application

Term Extract¹², TesauroVai¹³ and SketchEngine¹⁴) and open source (such as TermSuite¹⁵, TermoStat Web¹⁶ and FiveFilters¹⁷). 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.

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 platform¹⁸ [14], the extracted terms can be enriched with candidate translations obtained from external resources; SketchEngine¹⁹ [15] identifies collocates for the extracted terms from the source corpus; PoolParty²⁰ [16] allows the manual creation of hierarchies and the manual linking to resources such as DBpedia²¹; Saffron²² [17] suggests hierarchical relations between terms, to be afterwards supervised, and VocBench²³ [18] [19] allows the collaborative manual edition of vocabularies.

With regard to semi-automatic terminology enrichment, we also find several approaches in the litera-

ture. In [20], the enrichment consists in adding terms to a source thesaurus by exploiting parallel corpora. In [21], WordNet is used to establish *hierarchical* relations between the source terms. Oliveira and Gomes [22] 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 [23]. 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. In combination with Wikipedia and other resources, it conforms 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 this kind is the RDF dump of IATE, an effort described

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/>

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/>

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*²⁴, which gathers data from 21 different domains, being half of them closely related to legal activities. Several scientific works have been devoted to the conversion of *EuroVoc* into RDF [30–32] and it is now publicly available through an 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 [34], both of them modelled according to SKOS²⁵. Unlike *EuroVoc*, their content is focused on one specific domain, and can be of a 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 biligual (English-German). The same issue concerns the *UNESCO Thesaurus*²⁶, which provides multidisciplinary terminology in English, French, Spanish and Russian. Finally, the *International Labour Organisation Thesaurus*²⁷ 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 termi-

nology 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 corpus 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 [35], Socioterminology [36], Sociocognitive Theory of Terminology [37] or Frame-based Theory [38]). 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 [39], 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). 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 function as 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 terminologist in this step, several authors propose guidelines in the form of criteria that lexical units have to satisfy to be considered terms [35] [40]. Then, the meaning of a unit is to be discovered in text, and to be constructed through the relations that are specified 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 take the form of an ontology, as suggested in [38], and is the approach taken

²⁴<https://publications.europa.eu/en/web/eu-vocabularies>

²⁵<https://www.w3.org/TR/skos-reference/>

²⁶<http://vocabularies.unesco.org/browser/thesaurus>

²⁷<https://metadata.ilo.org/thesaurus/>

1 by the so called *terminological knowledge bases*, as
2 dubbed in [41], in which a knowledge base compo-
3 nent is enriched with a linguistic (terminological) com-
4 ponent. Some well-known examples of terminological
5 knowledge bases in different areas are GENOMA-KB
6 [42], OncoTerm [43] or EcoLexicon [44] [45].

7 These theories also propound that terms are to be
8 analysed as used in real communication by experts in
9 the domain, and that this may result in identifying var-
10 ious forms of designations (synonyms or term vari-
11 ants). Variants are to be accounted for in terminolog-
12 ical resources, as well as the causes for that varia-
13 tion [46]. Depending on the purpose of the resource at
14 hand, additional linguistic descriptions are also com-
15 mon in terminological resources, namely, source of the
16 term, morphosyntactic information, definition, refer-
17 ences to other terms (which can be of different na-
18 ture, e.g. synonyms, hyponyms, antonyms), usage con-
19 texts (that show how the term behaves in real texts), us-
20 age notes, or phraseology. Terms are usually assigned
21 to a domain, and all sources from which information
22 has been obtained are referenced, together with other
23 metadata (author, date, reliability degree, etc.).

24 When considering the multilingual perspective, best
25 practices in terminology work recommend that equiv-
26 alents in other languages are also collected from
27 domain-specific corpus in the languages of interest, as
28 well as the rest of linguistic descriptions [35]. An exact
29 equivalence relation is assumed when terms in mul-
30 tiple languages are related to a source term, although
31 language and culture differences may be captured in
32 the form of notes. However, previous works on mul-
33 tilingual terminological knowledge bases in the legal
34 domain show how important it is to explicit culture-
35 specific knowledge as intermediate representations as-
36 sociated with a common shared ontology [47].

37 Finally, we briefly refer to the theoretical stud-
38 ies (and practical applications) made by terminolo-
39 gists about terminological or conceptual relationships
40 between terms. Conscious of the importance of ac-
41 counting for such relationships in term banks, ter-
42 minographers have characterised them, studied them
43 in particular domains, and created methods for iden-
44 tifying them in corpora. The most important rela-
45 tions in this regard are the so-called hierachical re-
46 lationships (hyperonymy-hyponymy and meronymy).
47 However, several non-hierarchical relationships have
48 been intensively studied in some particular domains
49 (cause-effect, entity-function), and others have also
50 been considered for inclusion in terminological re-
51 sources (antonymy, synonymy, derivative relationships,

co-occurents and collocations). For a nice overview
we refer the interested reader to [48].

4. Requirements

7 The development of the first version of TermitUp
8 was guided by a set of requirements derived from
9 the study of existing language technologies, specifi-
10 cally those that deal with terminology, and the obser-
11 vation of their results, as well as from numerous dis-
12 cussions between the linguists, computer scientists and
13 researchers involved in this project.

14 **R1. Enrichment.** When confronted with domain
15 specific data, there is a need for identifying the specific
16 terms used in texts, as well as their meaning. Plain lists
17 of terms tend not to suffice if they are to be used for
18 annotation, classification or disambiguation and other
19 complex NLP tasks. Definitions, morpho-syntactic in-
20 formation, term variants and explicit relations amongst
21 terms can contribute significantly to improve the per-
22 formance of subsequent text processing tasks.

23 **R2. Multilingualism.** As already mentioned, inter-
24 national institutions have catered for the creation of
25 multilingual terminologies or thesauri to meet their
26 needs. However, these do not necessarily cover the
27 needs of a company or project in terms of languages,
28 or the purposes of the system being developed. This re-
29 sults in the need for systems that assist in the creation
30 of ad-hoc terminologies for certain language combina-
31 tions. There have been some initial attempts to develop
32 terminology extractors that work on multilingual cor-
33 pora, but results are still preliminary.

34 **R3. Disambiguation.** Although traditional theories
35 to terminology and language planning have backed the
36 approach that the terms in a domain are unambiguous,
37 unique and semantically precise, corpus-based termi-
38 nology studies have demonstrated that term variation
39 or synonym is common also in domain specific areas,
40 and that texts may also vary in the degree of specificiy.
41 Additionally, external language resources reused (see
42 Requirement 4) may contain different senses of a term,
43 since they are usually of a general character rather than
44 domain specific. This is translated into a necessity for a
45 disambiguation step when matching corpus-extracted
46 terms with external ones.

47 **R4. Reusability and Standardisation.** Knowledge
48 reuse is the cornerstone of Linked Open Data [4] and
49 the main goal of TermitUp. To meet this objective, this
50 service extracts knowledge from existing resources in
51 the *LLOD* cloud and publishes the resulting terminolo-

gies 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 service²⁸ [49] [50]. 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 Freeling²⁹. However, the performance of the tagger in a preliminary study was not satisfactory compared to other state-of-the-art part-of-speech taggers for Spanish. 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 aroused 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 repository³⁰.

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 [35] [51] [52].

Traditionally, nouns were considered the main parts of speech to be included in terminological resources [53], 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 patterns 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 Añotador³¹ [54], a service to identify dates and temporal expressions, so that we could also remove them, together with some additional noisy elements.

²⁸<https://sourceforge.net/projects/tbxtools/files/>

²⁹<http://nlp.lsi.upc.edu/freeling/>

³⁰<https://github.com/Pret-a-LLOD/termitup/tree/master/data>

³¹<https://annotador.oeg.fi.upm.es/>

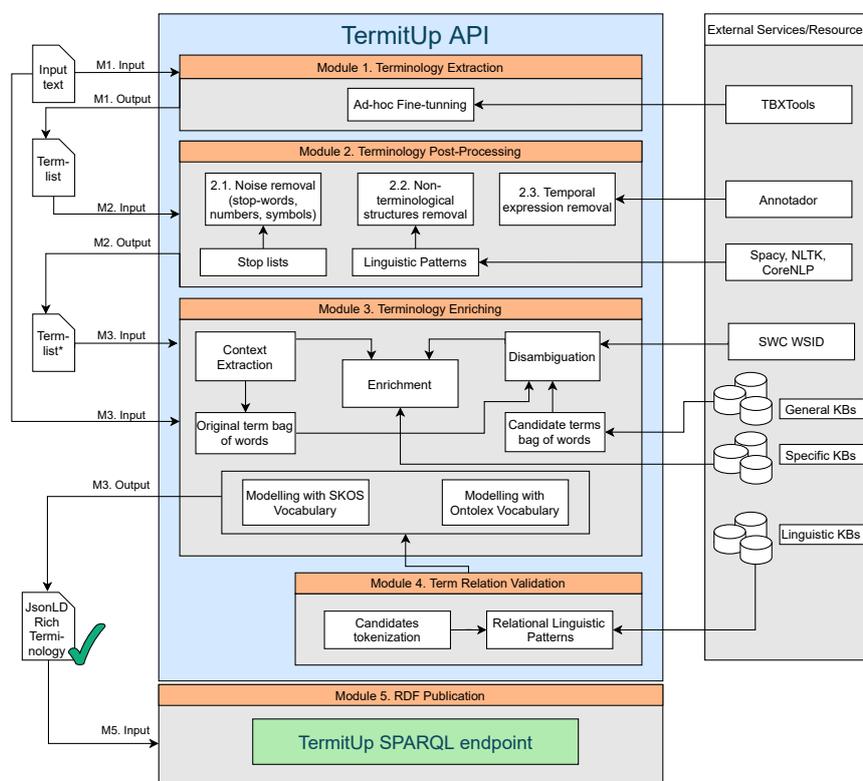


Fig. 2. TermitUp Architecture

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	Temporal Expressions in Spanish
[ADV]	simultáneamente (simultaneously)	12 de febrero (February 12th)
[ADV] + [ADJ]	inmediatamente posteriores (immediately after)	diez semanas (ten weeks)
[ADJ] + [ADV]	ininterrumpida inmediatamente (uninterrupted immediately)	quince días (fifteen days)
[NOUN] + [AUX]	partes deberán (parts shall)	nueve meses (nine months)
[NOUN] + [VERB]	consultas corresponderá (enquiries will correspond)	febrero de 2012 (February 2012)
[VERB] + [ADJ]	quedar constituida (be established)	meses siguientes (following months)
[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)	

5.3. Module 3: Terminology Enriching

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) algorithm³² based on BERT³³.

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 ($Ct1...Ct5$). The candidate target sense indicators ($s1...sn$) 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 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 en-

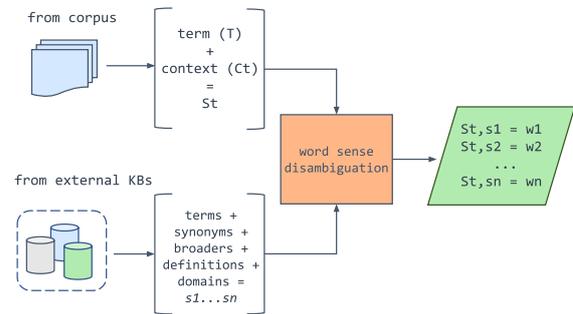


Fig. 3. Representation of the Word Sense Disambiguation workflow

rich 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: Enriching; 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 obtained, for instance, from Wikidata³⁴. The aim of this module is, thus, to check if such relations are correct.

This approach is inspired by the X-bar theory, that states that the formation of multi-word terms follows a hierarchical structure [35]. The approach suggests a comparison amongst the tokens of terms $t1$ and $t2$, and the token synonyms $s1t1...snt1$ and $s1t2...snt2$ 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 ConceptNet³⁵, 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).

³²https://github.com/semantic-web-company/ptlm_wsdl

³³<https://github.com/google-research/bert>

³⁴<https://www.wikidata.org/>

³⁵<http://conceptnet.io/>

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			
Ct1	the difficulty of retaining the hearing date arising from the practical difficulties for the witness	s1	s2	s3	
Ct2	after consideration on the papers by Her Honour Judge Stacey, the ET hearing has since been postponed	Ct1	4.45	6.10	5.58
Ct3	it seems that there had been an early case management hearing on 10 April 2017	Ct2	7.44	7.46	7.02
Ct4	the Tribunal may order any person in Great Britain to attend a hearing to give evidence	Ct3	6.22	7.79	6.88
Ct5	an application for a witness order may be made at a hearing or by an application in writing to the Tribunal	Ct4	7.17	7.94	7.82
Ct5		Ct5	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, which RDF version is limited and outdated, and its JSON API offers more complete and up-to-date data.

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

Additionally, we have implemented a set of rules based on post-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 application of the terminologies, one model will be more suitable than other. For example, if the user wants to use this terminology with a tool designed to specifically manage taxonomies, such as PoolParty or

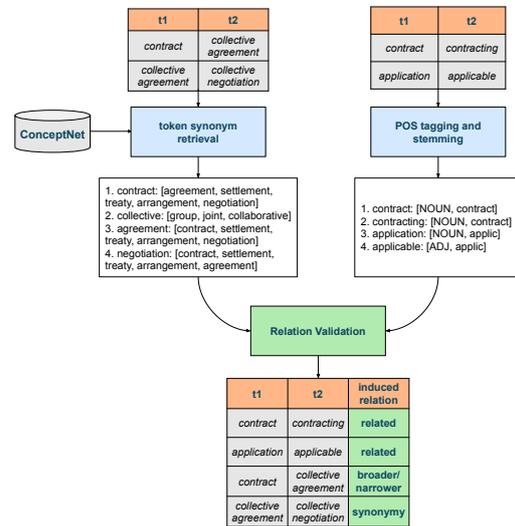


Fig. 4. Relation Validation Process

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 will be the most appropriate. Fig-

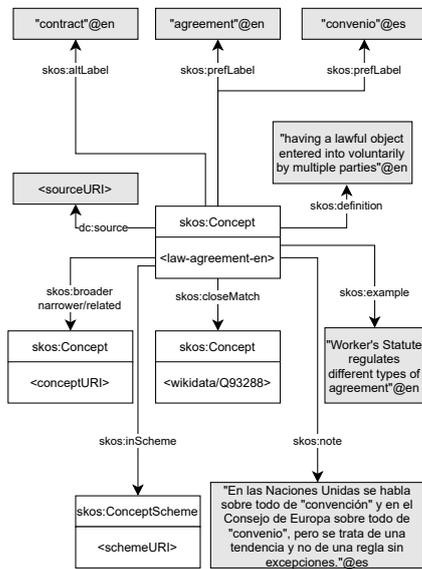


Fig. 5. Example of term modelled with SKOS

ures 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 in 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 Editor³⁶ that can be subsequently accessed and queried by the user. 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-LLOD³⁷ 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.

³⁶<https://termitup.oeg.fi.upm.es/sparql>

³⁷<https://pret-a-llod.eu/>

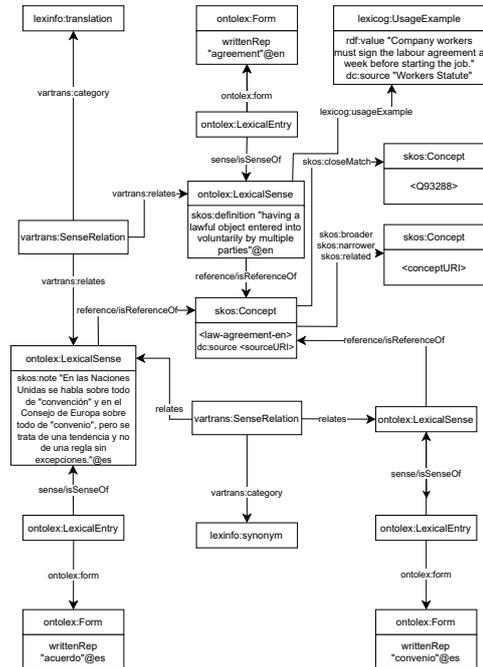


Fig. 6. Example of term modelled with Ontolex

Much as Prêt-à-LLOD is demonstrated in three pilots of disparate nature, spanning radically different domains such as the pharmaceutical and the e-government ones, TermitUp can be used in a number of different contexts. TermitUp has been applied to produce a terminology of labor law in the framework of the H2020 Lynx project³⁸ 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 platform³⁹ or downloadable as a static bulk dataset in Zenodo⁴⁰. In the context of the project supported by Grupo CPOnet⁴¹, 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 very different language registers.

³⁸<https://lynx-project.eu/>

³⁹<http://lkg.lynx-project.eu/kos>

⁴⁰<https://zenodo.org/communities/lynx/?page=1&size=20>

⁴¹<https://www.grupocponet.com/>

1 But the impact of TermitUp goes beyond these
2 domain-specific applications. Its use as a streamlined
3 component in composite workflows suggests a wider
4 range of applications. TermitUp might be used to cre-
5 ate user-specific terminologies, contribute to the lin-
6 guistic analysis of a community, or create more precise
7 vector models, with new features corresponding to the
8 links discovered by TermitUp. In its latest application
9 within the SmarTerp project⁴², TermitUp-craft termi-
10 nologies will support interpreting professionals by pro-
11 viding them extra information on the discourse at real
12 time –with a number of challenges related to efficiency
13 pending to be solved, since the project just started.

14 TermitUp is available in a public GitHub reposi-
15 tory⁴³, as a Python projec licensed under Apache Li-
16 cense 2.0 terms. The functionality is also available
17 through a HTTP REST API, thus satisfying Require-
18 ment 7. These web services are described using Ope-
19 nAPI⁴⁵, and they are running in web servers supported
20 by the Prêt-à-LLOD project. A stable release of the
21 software has also been published in the Zenodo plat-
22 form⁴⁶, favoring the preservation and reproducibility
23 of the research work.

24 7. Discussion

25
26
27
28 The main limitation found during the development
29 of this service is related to the publication of enriched
30 terminologies in RDF, i.e., to Requirement 5. The ob-
31 jective of this requirement is to maintain the traceabil-
32 ity of the data, since the provenance of the information
33 is an essential indicator of its quality. Thus, TermitUp
34 endeavours to store all sources of the collected data.

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

- 38 – *Terms, synonyms and translations*: In SKOS, they
39 are treated as literals, represented with the prop-
40 erties `skos:prefLabel` and `skos:altLabel`, that do
41 not allow to attach any additional information to
42 them. SKOS-XL⁴⁷, on the other hand, extends the
43 model to treat these properties as classes, being
44 able to preserve the source. In Ontolex, terms,
45
46

1 synonyms and translations are represented as
2 classes, allowing the representation of its source.

- 3 – *Context*: the context of a term is treated as an ex-
4 ample of how it is used within a text. Therefore,
5 the most suitable property to represent it in SKOS
6 is `skos:example` (subproperty of `skos:note`⁴⁸), that
7 allows representing text strings but no additional
8 information. In Ontolex, on the other hand, the
9 Lexicography module [55] considers this need
10 and introduces the `lexicog:UsageExample` class,
11 that on the contrary, allows representing more in-
12 formation beyond the text itself.
- 13 – *Term note*: this is a key element of traditional ter-
14 minology cards that provides additional informa-
15 tion, such as usage recommendations and domain
16 data. Some of the modern language resources do
17 not use term notes anymore, but others still keep
18 them, thus, we consider them valuable pieces of
19 knowledge for language professionals that need
20 to be preserved. In SKOS, term notes can be
21 modeled with `skos:note` and in Ontolex with `on-
22 tolex:usage`, both object properties pointing to lit-
23 erals. This implies that if we collect term notes
24 from different language resources, we would not
25 be able to model their provenance.
- 26 – *Definitions*: the same occurs with definitions,
27 since SKOS vocabulary applies `skos:definition`,
28 that is also a subproperty of `skos:note`, therefore
29 an object property that points to a literal. Ontolex
30 does not propose any class for definitions either,
31 and also employs `skos:definition`. We therefore
32 have the same issue to model its provenance.

33 Besides the difficulties stated above, we face an-
34 other modelling decision, since we find different types
35 of sources at different levels. This is, the language
36 resources with which the terms are enriched (i.e.
37 IATE) can be understood as *intermediate sources*,
38 that could be represented with the `schema:provider`
39 property⁴⁹. Intermediate sources are different from
40 *original sources*, that could be either a corpus (i.e.
41 European Legislation), an organisation (i.e. Euro-
42 pean Commission), an application (i.e. Definition Ex-
43 tractor) and even, or an individual (i.e. John Doe,
44 European terminologist). For their RDF representa-
45 tion, we consider properties such as `dc:source` and
46 `dct:BibliographicResource` from DublinCore⁵⁰ and
47

48 ⁴²[https://kunveno.digital/our-proyect/
4344](https://kunveno.digital/our-proyect/4344)

49 ⁴⁵<http://termitup.oeg.fi.upm.es/swagger>

50 ⁴⁶<https://doi.org/10.5281/zenodo.4461806>

51 ⁴⁷<https://www.w3.org/TR/skos-reference/#xl>

48 ⁴⁸<https://www.w3.org/TR/skos-reference#notes>

49 ⁴⁹<https://schema.org/provider>

50 ⁵⁰[https://www.dublincore.org/specifications/dublin-core/dcmi-
51 terms/](https://www.dublincore.org/specifications/dublin-core/dcmi-terms/)

1 the classes prov:Entity, prov:Agent, prov:Person and
2 prov:Organization from PROV ontology⁵¹.

3 Another discussion that aroused from the modelling
4 stage debates was whether the skos:definition (and re-
5 lated documentation properties) should be attached ei-
6 ther to the skos:Concept or to the skos:LexicalSense.
7 The SKOS specification remains vague in this point,
8 and both approaches are at least syntactically sound –
9 neither skos:definition nor its superproperty skos:note
10 declare a rdfs:domain. This freedom suggests a flexible
11 use which might be suitable to capture some subtleties.

12 First, when terminological data is transformed from
13 different sources, definitions sometimes seem attached
14 to concepts (e.g. data imported from Wikidata quali-
15 fies concepts), sometimes lexical senses (e.g. data im-
16 ported from WordNet). We suggest the application of
17 skos:definition in a flexible manner, being its subject a
18 skos:Concept or a lemon:LexicalSense at discretion.

19 Second, this loosen specification brings about the
20 opportunity to distinguish reference and sense, in
21 *fregean* terms. In his famous essay *Über Sinn und Be-*
22 *deutung* (1892), Gottlob Frege told apart the reference
23 and the sense of expressions [56]. In this writing, Frege
24 uses the example of Venus: both "the morning star" and
25 "the evening star" refer to the same object, Venus, but
26 the thought they express is rather different. The sense
27 is a mode of presentation, illuminating only a single
28 aspect of the referent. We wonder whether computers
29 can capture these nuances. We can certainly make such
30 an effort, reserving the objective information about
31 Venus for its skos:Concept (e.g. radius = 3000 km),
32 but administer the different subjective perceptions the
33 different components of the synset. Perhaps we want
34 to attribute the lemon:LexicalSense "Venus" a rela-
35 tively neutral subjective value related to celestial bod-
36 ies, and give the lemon:LexicalSense "morning star" a
37 hotter affective valence, possibly related to a more po-
38 etic context. These definitions and affective valences
39 will be necessarily stereotypes, not reflecting objec-
40 tive values (which are different for each mind), but in-
41 tersubjective, namely, reflecting common perceptions
42 and images (we refer the reader to [57] for more infor-
43 mation about emotional words).

44 We wonder whether personalized lemonizations will
45 ever be possible, describing the linguistic realities of
46 specific individuals, perhaps inferred from personal
47 big data such as personal email inboxes or alike.
48 But this endeavour is well beyond the scope of this
49

50
51 ⁵¹<https://www.w3.org/TR/prov-o/>

1 paper –we only stress the opportunity of attributing
2 skos:definition (and other triples) to skos:Concept or
3 lemon:LexicalSense in the most beneficial manner.

4 We have therefore gathered such ongoing discus-
5 sions on modelling issues in a proposal for good prac-
6 tices to model terminological resources, published as a
7 *Terminology* draft in the wiki of the Ontology-lexicon
8 Community Group in the W3C⁵², where we expose
9 background, motivation and use cases, and suggest
10 complementary elements to the existing models. Such
11 modelling modifications, naturally, need to be agreed
12 by the community before applying them.

13 8. Conclusions and Future Work

14
15
16
17 The automation in the generation of language re-
18 sources (specifically, terminological resources) is a
19 challenging task still unresolved. Automatic terminol-
20 ogy extraction and terminology management tools pro-
21 vide a good starting point and excellent assistance both
22 for terminology experts and language professionals,
23 but substantial manual effort is still required.

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

36 TermitUp helps covering those gaps by extracting
37 and post-processing terms from domain specific cor-
38 pora, and enriching them with translations, synonyms,
39 definitions, usage notes and terminological relations.
40 Consequently, this application establishes links to the
41 resources exploited, contributing to the population of
42 the *LLOD* with domain expert knowledge. Addition-
43 ally, the tool offers a module that helps validate the
44 terminological relations retrieved, that sometimes may
45 be imprecise. Finally, the tool structures the resulting
46 enriched terminologies, either following SKOS or On-
47 tolex model; and stores them in a Virtuoso SPARQL
48 Editor so that they can be freely accessed.
49

50
51 ⁵²<https://www.w3.org/community/ontolex/wiki/Terminology>

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 [58], 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 [59], 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 terminologi-

cal 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.

Acknowledgements

This work has been supported by the European Union's Horizon 2020 research and innovation programme through the Prêt-à-LLOD⁵³ 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* (2014).
- [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*, 2012.
- [3] J. Flisar and V. Podgorelec, Improving Short Text Classification using Information from DBpedia Ontology, *Fundamenta Informaticae* (2020).
- [4] T. Berners-Lee, Linked data-design issues (2006), URL <http://www.w3.org/DesignIssues/LinkedData.html> (2011).
- [5] J. Vivaldi, I. Da Cunha, J.-M. Torres-Moreno and P. Velázquez-Morales, Automatic Summarization Using Terminological and Semantic Resources., in: *LREC*, 2010.
- [6] K. Simon and K. Tomanek, EuroVoc-Based Summarization of European Case Law, in: *AI Approaches to the Complexity of Legal Systems: AICOL International Workshops 2015-2017*, Springer, 2018.
- [7] M. Arcan, M. Turchi, S. Tonelli and P. Buitelaar, Leveraging bilingual terminology to improve machine translation in a CAT environment (2017).
- [8] D. Tiscornia, The LOIS project: Lexical ontologies for legal information sharing, in: *Proceedings of the V Legislative XML Workshop*, Citeseer, 2006.
- [9] V. Lyding, E. Chiochetti, G. Sérasset and F. Brunet-Manquat, The LexALP information system: Term bank and corpus for multilingual legal terminology consolidated, 2006.
- [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* (2016).

⁵³<https://pret-a-llood.eu/>

- [11] S. Rose, D. Engel, N. Cramer and W. Cowley, Automatic keyword extraction from individual documents, *Text mining: applications and theory* (2010).
- [12] Z. Zhang, J. Gao and F. Ciravegna, Jate 2.0: Java automatic term extraction with apache solr, in: *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16)*, 2016.
- [13] M. Vázquez and A. Oliver, Improving term candidates selection using terminological tokens, *Terminology. International Journal of Theoretical and Applied Issues in Specialized Communication* (2018).
- [14] T. Gornostay, Terminology management in real use, in: *Proceedings of the 5th International Conference Applied Linguistics in Science and Education*, 2010.
- [15] A. Kilgarriff, P. Rychly, P. Smrz and D. Tugwell, Itri-04-08 the sketch engine, *Information Technology* (2004).
- [16] T. Schandl and A. Blumauer, PoolParty: SKOS thesaurus management utilizing linked data, in: *Extended Semantic Web Conference*, Springer, 2010.
- [17] 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.
- [18] 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.
- [19] 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* (2017).
- [20] H. Déjean, E. Gaussier, J.-M. Renders and F. Sadat, Automatic processing of multilingual medical terminology: applications to thesaurus enrichment and cross-language information retrieval, *Artificial Intelligence in Medicine* (2005).
- [21] L. Hollink, V. Malaisé and G. Schreiber, Thesaurus enrichment for query expansion in audiovisual archives, *Multimedia Tools and Applications* (2010).
- [22] H.G. Oliveira and P. Gomes, Towards the automatic enrichment of a thesaurus with information in dictionaries, *Expert Systems* (2013).
- [23] Y. Wu, Enriching a thesaurus as a better question-answering tool and information retrieval aid, *Journal of Information Science* (2018).
- [24] G.A. Miller, WordNet: a lexical database for English, *Communications of the ACM* (1995).
- [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*, Association for Computational Linguistics, 2010.
- [27] J. Gracia, M. Villegas, A. Gómez-Pérez and N. Bel, The apertium bilingual dictionaries on the web of data, *Semantic Web* (2017).
- [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 (2015).
- [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.
- [30] L.P. Paredes, J. Rodríguez 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).
- [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 (2009).
- [32] L.A. Díez, B. Pérez-León, M. Martínez-González and D.-J.V. Blanco, Propuesta de representación del tesoro Eurovoc en SKOS para su integración en sistemas de información jurídica, *Scire: representación y organización del conocimiento* (2010).
- [33] B. Zapilko, J. Schaible, P. Mayr and B. Mathiak, TheSoz: A SKOS representation of the thesaurus for the social sciences, *Semantic Web* (2013).
- [34] J. Neubert, Bringing the "Thesaurus for Economics" on to the Web of Linked Data., *LDOW* (2009).
- [35] M.T. CABRÉ, La terminología: teoría, metodología, aplicaciones. Barcelona: Editorial Antártida, *Empúries* (1993).
- [36] F. Gaudin, *Socioterminologie*, Publication Univ Rouen Havre, 1993.
- [37] R. Temmerman, *Towards new ways of terminology description: The sociocognitive-approach*, Vol. 3, John Benjamins Publishing, 2000.
- [38] P. Faber, Frames as a framework for terminology, *Handbook of terminology* (2015).
- [39] C. Barriere and A. Agbago, TerminWeb: a software environment for term study in rich contexts, in: *International Conference on Terminology, Standardisation and Technology Transfer (TSTT 2006)*, 2006.
- [40] M.-C. L'Homme, *Lexical Semantics for Terminology: An Introduction*, John Benjamins Publishing Company, 2020.
- [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.
- [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: *LREC*, 2004.
- [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).
- [44] P. Faber, P. León-Araúz and A. Reimerink, EcoLexicon: new features and challenges, *GLOBALEX* (2016).
- [45] A. San Martín, M.C. García and M.B. Castro, Presente y futuro de la base de conocimiento terminológica EcoLexicon, *Onomázein: Revista de lingüística, filología y traducción de la Pontificia Universidad Católica de Chile* (2020).

- [46] M.T. Cabré, J. Freixa, M. Lorente and C. Tebé, La terminología hoy: replanteamiento o diversificación, *Organon* **12** (1998).
- [47] 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.
- [48] 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).
- [49] T. Oliver and M. Vázquez, A Free Terminology Extraction Suite, in: *Proceedings of the Twenty-ninth International Conference on Translating and the Computer*, 2007.
- [50] A. Oliver and M. Vázquez, TBXTools: a free, fast and flexible tool for automatic terminology extraction, in: *Proceedings of the International Conference Recent Advances in Natural Language Processing*, 2015.
- [51] E. Alcaraz and B. Hughes, El español jurídico, *Barcelona: Ariel* (2002).
- [52] E. Alcaraz and B. Hughes, *Legal translation explained*, Routledge, 2014.
- [53] M.-C. L'Homme, What can verbs and adjectives tell us about terms?, *Terminology and Knowledge Engineering, TKE 2002. Proceedings* (2002).
- [54] M. Navas-Loro and V. Rodríguez-Doncel, Annotador: a Temporal Tagger for Spanish, *Journal of Intelligent and Fuzzy Systems* (2020).
- [55] J. Bosque-Gil, J. Gracia and E. Montiel-Ponsoda, Towards a Module for Lexicography in OntoLex., in: *LDK Workshops*, 2017.
- [56] G. Frege, Über sinn und bedeutung, *Zeitschrift für Philosophie und philosophische Kritik* (1892).
- [57] M.M. Bradley and P.J. Lang, Affective norms for English words (ANEW): Instruction manual and affective ratings, Technical Report, Technical report C-1, the center for research in psychophysiology . . . , 1999.
- [58] 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.
- [59] J. Bosque-Gil, 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* (2016).