Proceedings of the 12th Conference on Language Resources and Evaluation (LREC 2020), pages 5654–5659
Marseille, 11–16 May 2020

Defying Wikidata: Validation of Terminological Relations in the Web of Data

Patricia Martín-Chozas[1], Sina Ahmadi[2], Elena Montiel-Ponsoda[1]
[1]Ontology Engineering Group, Universidad Politécnica de Madrid
[2]Insight Center for Data Analytics, National University of Ireland Galway
[1]{pmchozas, emontiel}@fi.upm.es, [2]sina.ahmadi@insight-centre.org

Abstract

In this paper we present an approach to validate terminological data retrieved from open encyclopaedic knowledge bases. This need arises from the enrichment of automatically extracted terms with information from existing resources in the *Linguistic Linked Open Data* cloud. Specifically, the resource employed for this enrichment is WIKIDATA, since it is one of the biggest knowledge bases freely available within the Semantic Web. During the experiment, we noticed that certain RDF properties in the Knowledge Base did not contain the data they are intended to represent, but a different type of information. In this paper we propose an approach to validate the retrieved data based on four axioms that rely on two linguistic theories: the x-bar theory and the multidimensional theory of terminology. The validation process is supported by a second knowledge base specialised in linguistic data; in this case, CONCEPTNET. In our experiment, we validate terms from the legal domain in four languages: Dutch, English, German and Spanish. The final aim is to generate a set of sound and reliable terminological resources in RDF to contribute to the population of the Linguistic Linked Open Data cloud.

Keywords: linguistic linked open data, knowledge representation, terminological relations, resource population, multilingualism

1. Introduction

We are living what many people call the *Fourth Industrial Revolution* that is distinguished by the strengthening of Information and Communications Technology, and specifically the emergence of the *Knowledge Society* (Bindé, 2005). In this society, data and knowledge have become very valuable assets: generating, sharing and reusing data are common transactions. However, not every area of knowledge is undergoing such transformation at the same speed; this is the case of the *legal domain*.

In the last decade, law publishers, such as the European Union, joined the effort of publishing documents in XML-based structured formats (Caterina Lupo et al., 2007). Similarly, there have also been several non-official initiatives (Chalkidis et al., 2017) (Frosterus et al., 2013) (Rodríguez-Doncel et al., 2018) to publish legislation as per the Linked Data Principles (Bizer et al., 2009). However, what seems to be the definitive push in the exposure of *legal resources as data* in Europe is the *European Legislation Identifier* (ELI) [1], an initiative to harmonise the manner in which legislation is published. Every piece of legislation is identified by an HTTP URI and homogeneously described with a common minimum set of metadata elements supported by the ELI Ontology. Still, legal affairs have always been difficult to understand by non-experts due to its intricate and complex jargon. Scarce documentation available online hardly helps improve this situation as many linguistic resources from the legal domain are still published in physical formats and those that are online usually present a non-machine-readable structure (published as PDF).

If we take the *LLOD* cloud[2] as reference of structured language resources, we find domain specific assets (terminologies, thesauri, vocabularies and knowledge bases) such as the UMTHES[3] (environmental thesaurus), GBA[4] (geological thesaurus, SentiWS[5]) (sentimental analysis vocabulary), amongst others. Nonetheless, very few resources in this cloud represent knowledge from the legal domain. To enrich this gap, the work described here aims at creating rich multilingual linked terminological resources in the legal domain published in Semantic Web formats.

For this purpose, our efforts are devoted to create legal terminologies by automatically extracting terms from corpus and enriching them with information from the Linguistic Linked Open Data cloud. In this process, we have encountered several issues with the data retrieved, specially when acquiring synonyms for our source terms. Consequently, we propose a relation validation approach to enrich language resources with curated terminological relations from freely available knowledge bases. The experiment has been oriented to the legal domain, specifically, labour law and legislation, but we have also tested the approach with terms from the industry domain, so it can be applied to different fields. Additionally, since we are extracting knowledge from open knowledge bases, this process can be used to assess the status of these resources and suggest methods to curate their data and improve their quality.

The rest of the paper is structured as follows: Section 2. contains an overview of available language resources in structured formats (specifically in Semantic Web formats) and related work in linked terminologies. Section 3. refers to the motivation for this work. Section 4. presents the proposed approach. Section 5. describes the experiment to test our approach and analyses the results from both, a qualitative and quantitative perspective. Finally, section 6. concludes the paper and proposes the future work.

[1] `https://publications.europa.eu/en/web/eu-vocabularies/eli`

[2] `http://linguistic-lod.org/`

[3] `https://lod-cloud.net/dataset/umthes`

[4] `https://lod-cloud.net/dataset/geological-survey-of-austria-thesaurus`

[5] `https://lod-cloud.net/dataset/sentiws`

2. Related Work

Various efforts have been devoted to publishing linguistic resources following Semantic Web standards (mainly, the Resource Description Framework, RDF (Klyne and Carroll, 2006)) and to link them as per Linked Data Principles in the *LLOD* cloud, which is composed by *Domain-independent resources* and *Domain-dependent resources*.

Domain-independent resources are the biggest assets in the *LLOD* cloud. WordNet (Miller, 1995), for instance, is a well known general lexicon of the English language that has been converted into RDF following the Lemon[6] model (McCrae et al., 2014) and linked with many other resources within the cloud. Its most valuable feature is the representation of different senses per lexical entry to avoid ambiguity issues. BabelNet is one of the resources that exploits the linked version of WordNet; in combination with Wikipedia it conforms a multilingual semantic network of encyclopedic and language content that covers several domains (Navigli and Ponzetto, 2012). This network maps the information retrieved from WordNet and Wikipedia through synsets with several senses. A similar work was developed in the transformation of the Apertium bilingual dictionaries into RDF, using the *lemon model* and specifically the `vartrans` module for translations (Gracia et al., 2017).

Concerning *domain-dependent resources*, one of the most notable contributions is the conversion of a data dump of *IATE*, the *(InterActive Terminology for Europe)*[7], a multilingual terminological database of the European Union. It was converted into RDF and linked with the European Migration Network glossary (Cimiano et al., 2015). In this case, the work relied on the *lemon* vocabulary to organise the lexical information and the *Simple Knowledge Organization System (SKOS)* to represent the concepts, since *SKOS* is a specific vocabulary to model taxonomies and thesauri. Another related work to *IATE* was devoted to enrich of its terminological entries with translations and contexts through Statistical Machine Translation (SMT) approaches (Arcan et al., 2018). This work was also published as RDF, using the PROV-O ontology (Lebo et al., 2013) to distinguish automatic data from those manually retrieved. A similar terminology conversion work was *Terminoteca RDF*, one of the most important projects in Spain in this field (Bosque-Gil et al., 2016b). Here, the work was focused in converting several multilingual Spanish terminologies in XML (*Terminesp* and *Termcat* glossaries) into RDF according to the Ontolex-lemon model[8], and publishing them as Linked Data (Bosque-Gil et al., 2016a).

Another significant work with regard to legal linguistic resources is the conversion into RDF of *EuroVoc*[9], the multilingual and multidisciplinary thesaurus of the European Union. EuroVoc is one of the most important resources in this field (Díez et al., 2010) and it has been entirely represented in *SKOS*. The European Union have also developed the *CELLAR* common repository of metadata and content, in which legislation is represented as linked data and can be accessed through a SPARQL endpoint[10] (Francesconi et al., 2015). Finally, the *STW Thesaurus for Economics* seems relevant to this work since it shares terminology with the legal domain and contains *broader/narrower* relations between its terms represented in *SKOS* (Neubert, 2009).

Despite these more targeted approaches to represent legal knowledge in the Web of Data, the information gathered is still very *general*, meaning that data about specific parts of law and different jurisdictions is not usually represented. Such need has already been analysed in previous works by the authors. A first approach to contribute to the *LLOD* cloud with legal data has been suggested in (Martín Chozas, 2018), but involves a considerable amount of manual work that hinders the management of massive amounts of data. Also, this previous approach did not work on improving the quality of the retrieved data as this approach proposes.

3. Motivation

This work is part of a wider project, Lynx[11], aimed at sharing legal knowledge amongst European citizens through a Multilingual Knowledge Graph. For a better access to the legal knowledge, corpora needs to be translated, annotated and classified. Consequently, we need legal language resources to support such tasks. We performed an extensive research on legal language resources and noticed that the most important legal assets for the language industry cannot be understood by machines, since they are not available in open and structured formats. A good example is the *Black's Law Dictionary*, a monolingual legal dictionary that contains links between entries, which enriches the value of the resource; however, it is presented as a physical dictionary.

Thus, we opted for the generation of our own legal terminologies through terminology extraction from corpora. Afterwards, given a list of domain-specific terms, we started populating each term with additional information by querying an encyclopaedic knowledge base (EKB), in our case, WIKIDATA[12]. Although we are aware of similar EKBs, particularly BABELNET (Navigli and Ponzetto, 2010) and DBPEDIA[13], we did not integrate them in the current study as the first one has a limitation in the number of the queries that can be daily executed and the latter has been shown in a comparative study (Abián et al., 2017) to offer lesser data with lower objectivity than WIKIDATA.

Since our goal is the retrieval of specific knowledge from a general resource, we need to filter ambiguous information in order to get data only from our domain of interest. To remedy this, we implemented a straightforward but effective technique based on the semantic relation of the retrieved instances and a set of ground-truth entities from the legal domain. The ground-truth entities were manually collected from the same EKB, i.e. WIKIDATA using the Wikidata Query Service[14], with the assumption that our target terms are semantically related to them through one of the following hiearchical relations: `part-of`, `superclass-of /`

[6] `https://lemon-model.net/`

[7] `http://iate.europa.eu`

[8] `https://www.w3.org/2016/05/ontolex/`

[9] `https://publications.europa.eu/en/web/eu-vocabularies`

[10] `http://data.europa.eu/euodp/data/dataset/sparql-cellar-of-the-publications-office`

[11] `http://lynx-project.eu/`

[12] `https://www.wikidata.org`

[13] `http://dbpedia.org`

[14] `https://query.wikidata.org/`

subclass-of and instance-of. Therefore, we started collecting definitions, translations and synonyms. The latter are specially valuable for tasks such as Question Answering, Query Expansion, Cross-lingual Search and Machine Translation. Synonyms in Wikidata are represented by the also known as property, usually intended to identify aliases, spelling variants, scientific names and nicknames[15]. However, during the evaluation of this linking experiment, we observed that, in many occasions, the also known as property in WIKIDATA (represented by the skos:altLabel property) did not describe a synonymy relation between terms but a different one, such as hypernymy, hyponymy or other non-hierarchical relations (example in Table 1).

Source Term	**Wikidata altLabel**	**Correct Relation**
"diskriminierung"@de	"soziale Diskriminierung"@de	*narrower*
"public policy"@en	"government policy"@en "state policy"@en	*related*
"Völkrrecht"@de	"allgemeines Völkerrecht"@de	*narrower*

Table 1: Examples of wrong altLabel relations

How, then, can we generate domain-specific resources by reusing existing knowledge bases, guaranteeing that the retrieved data are correct? Specifically:

- *How can we guarantee that we are retrieving synonymy and not another kind of terminological relation between our source and the target term?*

4. Proposed Approach

Our approach is based on two different linguistic theories that, combined, give birth to a series of axioms used to induce semantic relations. We analyse which kind of relation exists (if any) between source terms in our dataset and target terms in the Encyclopaedic Knowledge Base. In this way, we determine whether the also known as property describes a synonymy relation or if it presents another type of terminological relation: hypernymy (*broader terms*), hyponymy (*narrower terms*) or other (*related terms*).

To do so, our axioms are based on the X-bar theory stating that the formation of multiword terms follow a hierarchical structure as shown in Figure 1 (Cabré and Sager, 1993).

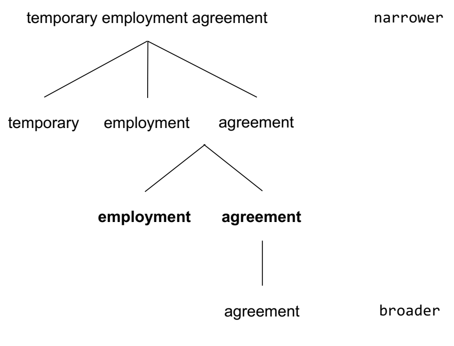

Figure 1: Multiword term as per the X-bar theory

[15] https://www.wikidata.org/wiki/Help:Aliases

The second theory studied follows a similar hierarchical approach for conceptual systems and multi-dimensionality of terms and also applies in our case that we are handling mono-dimensional or mono-hierarchical conceptual systems (Martínez et al., 2008) (Figure 2).

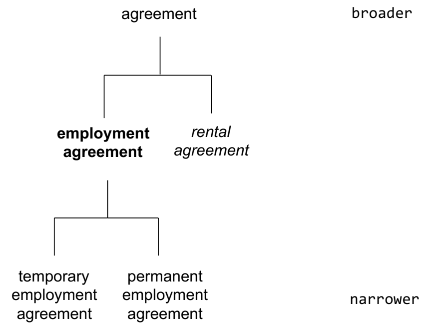

Figure 2: Multiword term as a monodimensional tree

Consequently, Table 2 the axioms that, based on these theories, can induce semantic relations. We perform a comparison between the tokens of the source term T, the tokens of alternative labels A and synonyms for the tokens of T and A that are retrieved from a linguistic knowledge base (see Figure 3). This approach is based on the assumption that the terms retrieved from the linguistic knowledge base (S) are consolidated synonyms of the source term T.

In this work, by token we refer to the consecutive characters delimited by spaces. Therefore, the tokens of T are shown as the set $\{t_1, t_2, ..., t_j, ...t_n\}$. In the axiom column, $\in!$ represents the unique existential quantifier, meaning *there exists exactly one* and S_{t_j} refers to the set of the synonyms retrieved for the token t_j. The key concept in determining synonymy, narrowness and broadness is the number of the tokens present in T and A defined as $|T|$ and $|A|$, respectively. On the other hand, relatedness is induced, meaning any kind of relation different from the three types previously mentioned. These axioms are iteratively applied over all the alternative labels. Examples corresponding to the axioms and the induced relations are provided in Table 2.

5. Experiment

Having a list of 112 law-case documents in PDF, published by the Government of the United Kingdom[16], we first converted the files into raw text and preprocessed them: normalising, encoding and removing noisy unstructured text. From this corpus, we used Tilde's[17] extraction services to obtain 300 random terms to test our approach.

In order to populate our initial plain term list with curated data, we extracted ground-truth entities in the legal domain by randomly selecting 20% of the terms and including the most generic entities associated to the terms with one of the following properties: part-of, superclass-of / subclass-of and instance-of. Some of these *legal subjects* are: *legal instrument (Q1428955)*, *common law (Q30216)*, *sociology of law (Q847034)*, *labour law (Q628967)*, *legal concept (Q2135465)* and *social action (Q769620).*

[16] https://www.legislation.gov.uk

[17] https://www.tilde.com/

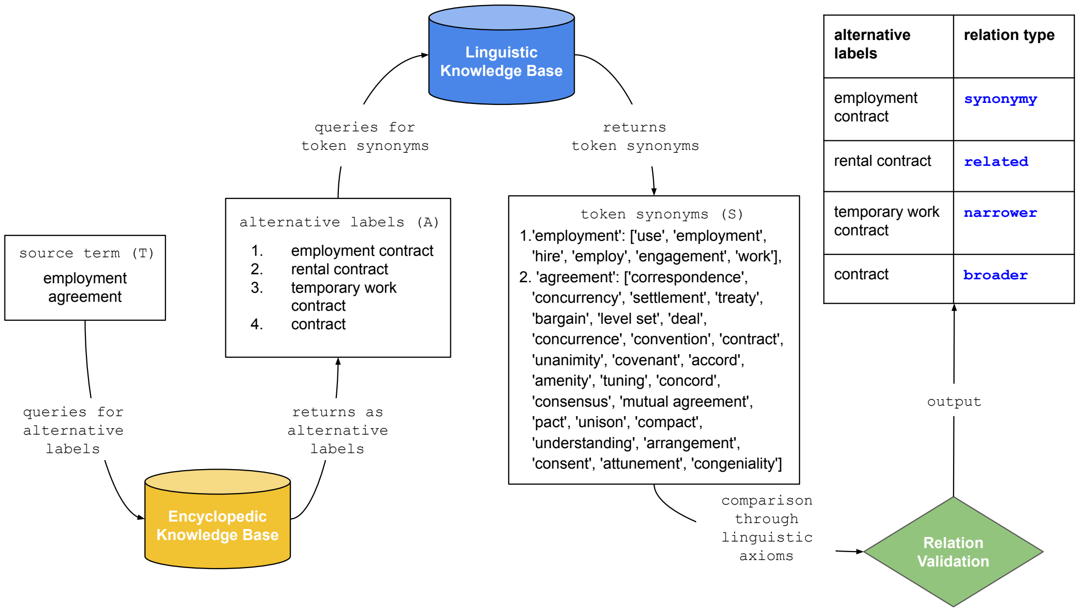

Figure 3: Example of the Relation Validation workflow

Axiom	Induction and Example
$(\lvert T\rvert = \lvert A\rvert) \wedge [\forall t_j \in T, \exists! a_i \in A, t_j = a_i \vee a_i \in S_{t_j}]$	**T and A are synonyms** T = "employment agreement" A = "employment contract" S = {(job, position, work), (contract, compromise, binding)}
$(\lvert T\rvert < \lvert A\rvert) \wedge [\forall t_j \in T, \exists a_i \in A, t_j = a_i \vee a_i \in S_{t_j}]$	**A is a narrower term of T** T = "employment agreement" A = "temporary work contract" S = {(job, position, work), (contract, compromise, binding)}
$(\lvert T\rvert > \lvert A\rvert) \wedge [\forall a_i \in A, \exists t_j \in T, a_i = t_j \vee a_i \in S_{t_j}]$	**A is a broader term of T** T = "employment agreement" A = "contract" S = {(job, position, work), (contract, compromise, binding)}
$[\exists t_j \in T, \exists a_i \in A, \exists s \in S_{t_j}, t_j = a_i \vee s \in A]$	**T and A are related** T = "employment agreement" A = "rental contract" S = {(job, position, work), (contract, compromise, binding)}

Table 2: Axioms for inducing semantic relations between alternative labels (A) of a term (T) using term synonyms (S)

From our initial list of legal terms, we had the 59% terms matching with WIKIDATA entries under the selected *legal subjects*. We realised that the organisation of WIKIDATA entries is heterogeneous and seems to be arbitrary. Several parent subjects such as *legal concept* or *legal profession* contain very few instances of narrower concepts under their scope: we discovered terms that were not related to any of these subjects and, consequently, could not be retrieved.

Then, we retrieve the matched concepts with five different types of data: translation equivalents, definitions, synonyms, broader terms and narrower terms. In this experiment, we have retrieved data in four languages (Dutch, English, German and Spanish), but the same workflow could be used to retrieve information for any of the other languages available in the queried knowledge base.

From the retrieved terms, only a 56% of the entries include data under the alternative label property (Figure 4). This means that a great part of the community efforts are devoted to the enrichment of the preferred label of WIKIDATA entries but not many to the alternative labels.

To validate the relations between the alternative labels and the main term label, we retrieved token synonyms from CONCEPTNET[18], from which 26 terms in Dutch, 27 terms in English, 38 terms in German and 21 terms in Spanish were available (Figure 5). These low numbers are not surprising due to the lack of open legal knowledge available on the web, as pointed out in the introduction.

Evaluating data retrieved from collaboratively-curated resources is a challenging task due to the completeness and ambiguity issues. To evaluate the relation validation performance, we created a gold-standard resource containing the manually corrected and completed retrieved data.

[18] http://conceptnet.io/

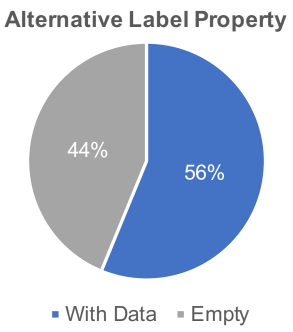

Figure 4: Wikidata alternative label (`altLabel`) status

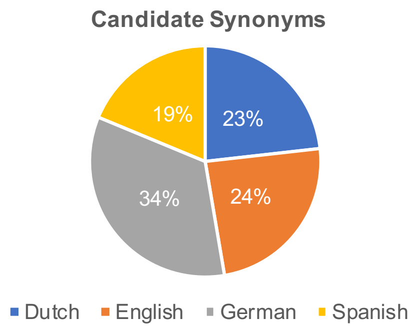

Figure 5: Candidate synonyms per language

We define *accuracy* as the number of the correctly induced relations with respect to all the relations.
Table 3 provides the evaluation results for the four languages, Dutch, English, German and Spanish.

Language	Matched	Missing	Accuracy
Dutch	26	12	78%
English	27	11	85.5%
German	38	19	75%
Spanish	21	11	72.5%

Table 3: Evaluation results and missing terms

The accuracy percentages show an acceptable performance of the approach. However, since for many of our terms we could not find any candidate synonym in CONCEPTNET we need to perform further evaluation with terms from a different domain. This allows us to check if the amount of synonyms retrieved from CONCEPTNET depends on the domain. Likewise, we can increase the retrieval of synonyms by querying additional linguistic knowledge bases, such as WORDNET or IATE. Finally, after the evaluation, we reorganised the candidates as per their correct relations with the source terms: synonyms, broader, narrower or related. To represent these data, we applied the SKOS vocabulary[19].

6. Conclusions and Future Work

Despite our initial hypotheses, based on the observations of Spanish terminologies, our experiments show that the performance is high also for terms in English, German and Dutch, at least for the legal domain. As a future work, our plan is to replicate the experiment with terminologies from the medical and the public administration domains.
Regarding the linking with WIKIDATA, the amount of matches retrieved is very low, meaning that legal information is barely represented within this resource. Publishing legal data in open machine-readable formats needs to be a common practice, specially in public organisations.
Our experiments also show that encyclopaedic or general knowledge bases usually contain data in their main properties: preferred labels and translations. However, additional data, such as definitions and alternative labels are rarely provided. This is why in the second process of this workflow, the amount of terms was significantly reduced (from 338 terms, 148 did not contain any alternative label). We stress the importance of developing terminological resources containing these type of data, specially synonyms, since they are highly valuable for several Natural Language Processing tasks, as mentioned in Section 3.
Similarly, linguistic knowledge bases do not contain synonyms for every term. This means that in many occasions, the axioms *would* have worked but they did not have any data to work with (78 cases without candidate synonyms). We need to perform more experiments with terms from different areas to check whether this issue depends on the domain or on the content of the queried knowledge base.
On the other hand, subject-disambiguation process is quite straightforward. This contribution does not focus on disambiguation techniques, but part of the future work is to test and implement state-of-the-art disambiguation methods.
Regarding our axioms, we want to increase their number of them to include relations such as part-whole, entity-function and generic-specific and semantic relations such as antonymy. The aim is to discover which kind of relations exist under the `skos:related` property. In addition, we want to test our axioms with more semantic resources: DBPEDIA, BABELNET, EUROVOC, IATE RDF and different WORDNETS. As all these resources are part of the *LLOD* cloud, adding them to the workflow is a way to enrich the legal knowledge gap in the Semantic Web.
On the whole, apart from validating the terminological information retrieved for our research purposes, testing the quality of the data in two huge knowledge bases as WIKIDATA and CONCEPTNET is quite a beneficial exercise for the community. Such experiments are necessary to continue enriching resources within the Semantic Web, and covering information needs of under-represented domains (such as the legal domain, in this case).
This work is openly available at `https://github.com/sinaahmadi/LDTerm`.

7. Acknowledgements

The authors would like to thank the three anonymous reviewers for their insightful suggestions and careful reading of the manuscript. This work has received funding from the EU's Horizon 2020 Research and Innovation programme through the contracts LYNX (grant agreement No. 780602), Prêt-à-LLOD (grant agreement No. 825182) and ELEXIS (grant agreement No. 731015).

[19] `https://www.w3.org/TR/skos-reference/`

8. Bibliographical References

Abián, D., Guerra, F., Martínez-Romanos, J., and Trillo-Lado, R. (2017). Wikidata and dbpedia: a comparative study. In *Semanitic Keyword-based Search on Structured Data Sources*, pages 142–154. Springer.

Arcan, M., Montiel-Ponsoda, E., McCrae, J. P., and Buitelaar, P. (2018). Automatic enrichment of terminological resources: the IATE RDF example. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC-2018)*.

Bindé, J. (2005). Towards knowledge societies: Unesco world report.

Bizer, C., Heath, T., and Berners-Lee, T. (2009). Linked data-the story so far. *International journal on semantic web and information systems*, 5(3):1–22.

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

Bosque-Gil, J., Montiel-Ponsoda, E., Gracia, J., and Aguado-de Cea, G. (2016b). Terminoteca RDF: a gathering point for multilingual terminologies in spain. *Term Bases and Linguistic Linked Open Data*, page 136.

Cabré, M. T. and Sager, J. (1993). *La terminología: teoría, metodología, aplicaciones*. Antártida/Empúries Barcelona.

Caterina Lupo, F. V., Enrico Francesconi, Monica Palmirani, R. W. E. d. M., and Alexander Boer, P. M. (2007). General xml format(s) for legal sources. *Estrella Deliverable 3.1*.

Chalkidis, I., Nikolaou, C., Soursos, P., and Koubarakis, M. (2017). Modeling and querying greek legislation using semantic web technologies. In *European Semantic Web Conference*, pages 591–606. Springer.

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

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

Francesconi, E., Küster, M. W., Gratz, P., and Thelen, S. (2015). The ontology-based approach of the publications office of the eu for document accessibility and open data services. In *International Conference on Electronic Government and the Information Systems Perspective*, pages 29–39. Springer.

Frosterus, M., Tuominen, J., Wahlroos, M., and Hyvönen, E. (2013). The finnish law as a linked data service. In *Extended Semantic Web Conference*, pages 289–290. Springer.

Gracia, J., Villegas, M., Gómez-Pérez, A., and Bel, N. (2017). The apertium bilingual dictionaries on the web of data. *Semantic Web*, (Preprint):1–10.

Klyne, G. and Carroll, J. J. (2006). Resource description framework (rdf): Concepts and abstract syntax.

Lebo, T., Sahoo, S., McGuinness, D., Belhajjame, K., Cheney, J., Corsar, D., Garijo, D., Soiland-Reyes, S., Zednik, S., and Zhao, J. (2013). Prov-o: The prov ontology. *W3C recommendation*.

Martín Chozas, P. (2018). *Towards a Linked Open Data Cloud of language resources in the legal domain*. UPM.

Martínez, S. M., Benítez, P. F., and Buendía-Castro, M. (2008). *Terminología para traductores e intérpretes: una perspectiva integrada*. Tragacanto.

McCrae, J., Fellbaum, C., and Cimiano, P. (2014). Publishing and linking wordnet using lemon and rdf. In *Proceedings of the 3rd Workshop on Linked Data in Linguistics*.

Miller, G. A. (1995). Wordnet: a lexical database for english. *Communications of the ACM*, 38(11):39–41.

Navigli, R. and Ponzetto, S. P. (2010). Babelnet: Building a very large multilingual semantic network. In *Proceedings of the 48th annual meeting of the association for computational linguistics*, pages 216–225. Association for Computational Linguistics.

Navigli, R. and Ponzetto, S. P. (2012). Babelnet: The automatic construction, evaluation and application of a wide-coverage multilingual semantic network. *Artificial Intelligence*, 193:217–250.

Neubert, J. (2009). Bringing the" thesaurus for economics" on to the web of linked data. *LDOW*, 25964.

Rodríguez-Doncel, V., Navas-Loro, M., Montiel-Ponsoda, E., and Casanovas, P. (2018). Spanish legislation as linked data. In *TERECOM@ JURIX*, pages 135–141.