Representing Terminological Data in the Semantic Web
A proposal based on OntoLex-lemon

Patricia Martín-Chozas[1]; Thierry Declerck[2]; Elena Montiel-Ponsoda[1]; Víctor Rodríguez-Doncel[1]

[1] Ontology Engineering Group, Universidad Politécnica de Madrid
[2] Deutsches Forschungszentrum für Künstliche Intelligenz

Abstract. This paper describes an approach to represent terminologies in the machine-readable format of the Semantic Web, which improves the interoperability between terminological resources and opens up new possibilities yet to be discovered. The study's motivation stems from the realization that the existing formalisms, such as SKOS or OntoLex-lemon, might not adequately capture the information within authoritative terminological resources. Therefore, we identified model requirements by formulating a set of Competency Questions derived from the analysis of terminological resources across various fields and domains, in line with the ontology development methodologies adopted in this work. During this analysis, we faced different representation challenges such as the various sources of term descriptions and the quality indicators related to terms. Consequently, we propose Termlex, a proposal based on the OntoLex-lemon model that combines the conceptual structure of the SKOS model with the lexical information as modelled in OntoLex-lemon. In Termlex, we define new classes and properties to cover the specific needs of terminological resources coming from a variety of approaches. The paper concludes with the instantiation of the Termlex model through three different use cases that follow different modelling approaches as a validation attempt.

Keywords: Terminology Modelling; Terminology Representation; Semantic Web; Linguistic Linked Data; OntoLex; Language Resources

1. Introduction

Terminologies, in the sense of resources that compile and describe the terms used in particular subject fields, are acknowledged to play a fundamental role in the management of specialized knowledge, especially in multilingual projects, with the aim of indexing corpora, translating domain specific documentation and performing cross-lingual queries, amongst other tasks (Hovenga 2022; Stefaniak 2017).

Traditionally, terminologies have been manually elaborated by domain experts, translators, and linguists in their efforts to describe specialized language and ensure the correct use of terms (Cabré 1993). However, depending on the creators' purposes, the potential uses of the resulting resources, and the theoretical approaches underpinning the creation process (also known as *terminography*), terminological resources may vary in type, quantity, and structure of the data (Fuertes-Oliveira and Tarp 2014). As for the type of terminological data commonly captured in a terminological resource, it seems to be strongly influenced by the purpose or intended use of the resource, as well as by the theoretical framework in which the resource has been conceived and the methodological

approach followed. Broadly speaking, two main approaches can be distinguished: the *onomasiological* and the *semasialogical* approach. As summarized in Temmerman (2000), the onomasiological approach, as understood by the Vienna School for Terminology, begins with concepts and attempts to define them as part of an extra-linguistic concept system. Terms are then introduced to designate such concepts. Contrary to this, the semasiological approach starts from the terms, as used in a specific corpus, and tries to account for their behavior in context, with the goal of identifying or creating a corresponding concept.

Based on our observations, terminological resources created from an onomasiological standpoint tend to include less lexical data than those resources that take a semasiological perspective. We have also observed that resources in the first group are widely used for standardization and normalization purposes and, on many occasions, result in controlled vocabularies. On the contrary, resources created from a semasiological perspective aim to account for the use of terms in domain-specialized corpora, and usually include usage examples, term variants, collocations, or pragmatic aspects of terms. Resulting terminologies attempt to cover the conceptual and linguistic needs of a wider range of users involved in a communication act (translators, interpreters, journalists, etc.). This is, however, not an official categorization, but a conclusion drafted from our experience with different terminology resources, and often the border between them is blurred.

From a computational perspective, the representation formats in which terminological resources are commonly available also vary substantially. Although this kind of resource was initially created for direct consultation by humans, in 1936, terminologists started to pursue a standardization process for the formalization of terminology and language resources, initiated with the predecessor of ISO, the International Federation of Standardizing Associations (ISA), under the ISA/TC 37 technical committee. In the context of ISO, this standardization work resumed in 1952 with the ISO TC/37 technical committee.[1] ISO proposes a standard for the formal representation of terminology databases, the Terminological Markup Framework or TMF (ISO 16642 2017)[2] and a standard for the creation of machine-readable lexical resources, the LMF (*Lexical Markup Framework*) (Francopoulo 2013). ISO 16642 2017 specifies a meta-model for terminology markup expressed in XML (a format designed to structure documents and data on the Web) with the aim of supporting the use and exchange of terminological data between computer systems.

With the aim of structuring and interchanging the knowledge represented in such terminological and related language resources, several standardization initiatives based on XML were more recently launched, such as the Text Encoding Initiative (TEI) (Ide and Véroni 1995) and, subsequently, the TermBase eXchange or TBX format (ISO 30042 2019)[3] (Melby 2015).

TBX is endorsed by ISO and by the Terminorgs (Terminology for Large Organizations) consortium.[4] It is claimed to be adopted by industry in their translation and localization projects. Due to the large amount of data categories that compose the TBX standard, a

[1] https://www.iso.org/committee/48104.html

[2] https://www.iso.org/standard/56063.html

[3] https://www.iso.org/standard/62510.html

[4] https://terminorgs.net/

series of public dialects[5] have been devised to help developers apply the model: TBX-Core, TBX-Min and TBX-Basic. TBX-Core contains the main elements of the standard, and it is mainly used to create new dialects, for instance, private dialects for companies, which must comply with the TBX-Core. TBX-Min is the simplest dialect, intended to create and store monolingual and bilingual glossaries. Finally, TBX-Basic is regarded as the main dialect for terminology exchange, since it is designed to efficiently store monolingual, bilingual and multilingual glossaries in straightforward XML. This is, therefore, the most used dialect in industry. More information about TBX dialects can be found in Melby (2012).

According to the TBX-Basic documentation, information items or data categories in TBX resources are grouped into three levels: concept, language, and term levels. The main items at the *concept level* are definition, subject field, and project. Information about the language in which information is provided, as well as definitions in several languages, are accounted for at the *language level*. Finally, any descriptions related to the term itself (term type, part of speech), usage status of the term or its actual use in a text (context) are part of the *term level*. As stated in the TBX-Basic documentation, there are only two mandatory data categories in a TBX resource: term and language. Definition, context, part of speech, and subject field are believed to disambiguate the term's meaning and are recommended to be included whenever possible.

These standardization initiatives seem to accommodate the needs of multilingual terminological projects in commercial production settings, in which terminology is fixed by the terminology management team at the early stages of product development. In this way, "consistent corporate language is prescribed to form an identity, establish clear communication, and thereby realize savings in time, resources and money".[6] These formats serve their purpose well in most cases. However, we claim that if terminologies were represented using the Semantic Web principles, there would be additional benefits, as summarized in the following.

Firstly, XML formats lack principled mechanisms to integrate information from several resources, for instance, when automatically creating terminologies extracting information from previously existing resources (see Section 5). One of the main advantages of Semantic Web standards is that they foster interoperability with other data sources regardless of the nature of the data (linguistic, conceptual, numerical, etc.), the origin, or the provider of the data.

Secondly, the XML hierarchy of traditional formats is less flexible than the graph-like RDF structure, e.g., a terminology can grow by merely adding information instead of having to create a new model (tree) from scratch.

Thirdly, TBX and similar standards fall short of proposing well-established methods to interlink linguistic (terminological) data available from several language resources. Here is where the Semantic Web can play a major role. There are mechanisms to model and express the relationships that exist between the classes of two different resources. In this

[5] https://www.tbxinfo.net/tbx-dialects/

[6] https://termcoord.eu/catalogue/e-books/terminology/terminology-for-large-organizations-terminology-starter-guide

way, the terms in two terminological resources can be stated to refer to the same concept, as will be further detailed in the next section.

1.1. Benefits of the Semantic Web

Back in the early 2000s, the *World Wide Web Consortium (W3C)* promoted the publication of data in structured, machine-readable, and interlinked formats, in which the meaning of language data is coded and can be interpreted by machines to achieve more complex and effective queries. This initiative is known as the *Semantic Web* or the *Web of Data*, whose main idea is that not only the documents are connected, but the information contained in these documents is also interlinked (Berners-Lee, Hendler and Lassila 2001).

The most common format for publishing data on the Semantic Web is the *Resource Description Framework (RDF)*, which supports the description of concepts, the representation of information, and the interchange of data on the web. The main element of this representation format is the *triple*, a subject-predicate-object structure in which the information is organized. RDF is at the core of the *Linked Open Data* paradigm for publishing information, based on the *Linked Data Principles* (Berners-Lee 2006) that, overall, state that resources need to be identified by a *Uniform Resource Identifier (URI)*, a unique identifier that follows the HTTP standard web protocols, and that resources need to contain pointers to other resources, also identified by URIs. In RDF, resources are to be understood as entities, things, and they can be of different nature (documents, physical objects, people, abstract concepts, data objects, etc.).

Concepts, words, and any relevant entity are represented by RDF resources and identified by URIs, whose meaning can be further refined by OWL *ontologies*. In Artificial Intelligence, an ontology has been traditionally defined as "an explicit specification of a conceptualization" (Gruber 1993, 1). In simpler terms, an ontology can be understood as a model or vocabulary to represent the concepts of a certain domain (Chandrasekaran, Josephson and Benjamins 1999). This usually implies the definition of *classes*, relations, and relations amongst those classes, called *properties*. To standardize their application, these properties are normally associated to a *domain* and a *range*, which represent the target and the source of the relation respectively.

Some of the most used ontologies are FOAF,[7] to describe people; DublinCore,[8] to describe metadata; or Schema,[9] to structure content in websites. An ontology is commonly referred to as a *vocabulary* or a *model*, usually depending on its complexity. Throughout this paper, we use these terms without distinction. For more ontologies, we refer to the Linked Open Vocabularies portal[10] (Vandenbussche et al. 2017) that allows for selecting the most adequate vocabulary depending on the domain requirements.

In the last decade, several projects have been launched to promote the publication of data in Linked Data formats. One of the most important initiatives is the *Linked Open Data project*, which pursued the publication of Linked Data under open licenses and that

[7] http://xmlns.com/foaf/spec/

[8] http://purl.org/dc/elements/1.1/

[9] http://schema.org/

[10] https://lov.linkeddata.es/

resulted in the creation of the *Linked Open Data cloud* (*LOD cloud* for short) (Bizer, Heath and Berners-Lee 2011) as the main source of Linked Data. The LOD cloud is divided into sub-clouds, such as the Geography cloud, the Governmental cloud, or the Biomedical cloud, and each of them is composed of interlinked datasets in that field. The most relevant cloud for our work is the *Linguistic Linked Open Data cloud (LLOD cloud)*, which contains language resources such as corpora, lexicons, dictionaries, terminologies, thesauri, knowledge bases and other linguistic data sources.[11]

The remaining sections are organized as follows: Section 2 describes the related work, including the existing models and their limitations; Section 3 identifies the representation challenges that arose from the analysis of a selection of terminological resources as case studies from which requirements for a common data model are extracted; Section 4 presents the Termlex proposal, including the methodology followed for the ontology development (Section 4.1) and the core classes and properties of the model (Section 4.2); Section 5 introduces three use cases in which Termlex is applied to cover terminological requirements of different resources and, finally, Section 6 presents the conclusions and identifies the main benefits of using this representation approach.

2. Related Work

The purpose of this section is twofold: on the one hand, to analyze the existing models to represent Linguistic Linked Data and, on the other, to identify efforts that publish terminological resources following those models.

The first steps towards the combination of terminologies and ontologies were introduced through the *termontography* concept by Rita Temmerman (Temmerman and Kerremans 2003), in which theories and methods for multilingual terminological analysis of the sociocognitive approach (Temmerman 2000) are combined with guidelines for ontological analysis. In the same line, years later the *ontoterminology* concept was proposed (Roche 2012), focused on publishing terminological resources whose conceptual systems are ontologies (Roche, Damas and Roche 2014).

On the technical side, one of the first initiatives to represent language resources in the Semantic Web was the OTR model (Reymonet, Thomas and Aussenac-Gilles 2007), conceived to represent terminologies in OWL format. This model was quite simple, mainly composed of the class Term that included languages and labels as properties. Similarly, the LIR (*Linguistic Information Repository*) was conceived to localize ontologies, providing the mechanisms to represent term equivalents and term types associated to ontology concepts (Montiel-Ponsoda et al. 2011).

Nonetheless, the most widely applied models to represent language resources in the Semantic Web are SKOS and OntoLex-lemon. The SKOS vocabulary was designed to express the basic structure and content of classification schemes and thesauri, but also of concept schemes embedded in glossaries and terminologies (Miles et al. 2005). It allows for the creation of hierarchies amongst terms and the representation of definitions, examples and notes. Later on, SKOS-XL was proposed, an extension of the SKOS vocabulary to represent labels as classes. Both SKOS and SKOS-XL have been widely applied for the representation of thesauri like the EuroVoc thesaurus, containing terms

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

belonging to domains under the European Union legislation (Díez-Alvite et al. 2010); the UNESCO thesaurus, intended for indexing and retrieval of UNESCO documentation (Sánchez 2016); the AgroVoc thesaurus, about the agricultural domain (Caracciolo et al. 2013); the TheSoz thesaurus of social sciences (Zapilko et al. 2013) or the STW thesaurus for economics.[12]

On the other hand, a more comprehensive model was later proposed to enrich ontologies with linguistic information: the *lemon* vocabulary (McCrae et al. 2012), which became the basis for a W3C Community Group, the Ontology-lexicon Community Group.[13] Under the auspices of this community group, lemon evolved into OntoLex-lemon with the same purpose, to provide a principled way to describe how "ontology entities, i.e. properties, classes, individuals, etc. can be realized in natural language".[14]

OntoLex-lemon is a concise and descriptive model that does not contain a complete collection of linguistic categories but relies on external vocabularies and ontologies, such as Lexinfo[15] or OLiA.[16] In its current status, since the official publication of the community report in May 2016, OntoLex-lemon consists of five modules, each one dedicated to certain types of linguistic descriptors. The original objective of this model was to *lexicalize ontologies,* but it has been employed in a great number of efforts to publish dictionaries as part of the Web of Data. This has been the case with the Apertium dictionaries (Gracia, Villegas and Gómez-Pérez 2018) and the multilingual global series of K Dictionaries (Bosque-Gil et al. 2016). It has also been used to represent terminologies, specifically those included in the TerminotecaRDF project (Bosque-Gil et al. 2016).

Additionally, the conversion of a subset of the InterActive Terminology of Europe (IATE) that was also linked to the European Migration Network glossary into the first lemon model by means of the TBX2RDF service (Cimiano et al. 2015) is also worth mentioning. This work combined the lemon vocabulary with complementary RDF properties from the TBX format and was further developed by the creation of Terme-à-LLOD, a platform to convert and host terminologies on TBX2RDF (Di Buono et al. 2020).

Still, after a review of available models to convert traditional electronic linguistic resources, specifically terminologies, into Semantic Web formats, we realized that certain information items contained in those resources could not be appropriately represented with existing RDF-based models, as explained in the next section. Consequently, we propose the Termlex model, which permits the full modelling of terminologies in the Semantic Web.

3. Representation Challenges: Analysis of Terminological Resources

Without claiming to be exhaustive, in this section, our aim is to examine some terminological resources generated following different approaches (onomasiological vs.

[12] http://zbw.eu/stw/version/latest/about

[13] https://www.w3.org/community/ontolex/

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

[15] https://www.lexinfo.net/

[16] http://www.acoli.informatik.uni-frankfurt.de/resources/olia/

semasiological), which are examples of widely used terminological resources since all of them are recommended by the Terminology site of the European Parliament Directorate-General for Translation.[17] This selection was made based on the discriminatory features of the resources, with the goal of analyzing those that have little in common so that the proposed model could accommodate the needs of different variants. Such discriminatory features are the domain, size, purpose, the type of data contained, and the theoretical approach followed in their creation.

According to this analysis, we identify unsolved representation needs from a Semantic Web and Linked Data perspective. To the best of our knowledge, there have been no previous attempts to provide an ontology to model the information typically contained in terminological resources with the ambition of being sufficiently general and flexible to accommodate the representation needs of a wide range of terminological resources. For this aim, we follow the recommendations given by the most adopted methodologies (Suárez-Figueroa, Gómez-Pérez and Villazón-Terrazas 2009) for the creation of ontologies, using Competency Questions, i.e., questions in natural language that account for the type of knowledge that the ontology should contain, and which represent the requirements that the ontology should fulfil (Uschold and Gruninger 1996).

Each of the terminological resources analyzed has been selected for a different reason but, overall, we believe they illustrate the diversity of existing terminological resources, which might respond to different needs and result from different approaches to terminological work. This is reflected in the structure, type and nature of the terminological data captured in them. The terminological resources selected for this analysis are presented in Table 1, together with some of their features.

The data shown in the examples from each of the resources arise from an iterative analysis of the resources:

1) We analyze the resource and identify the representation needs.
2) We explore if there are available vocabularies that cover these needs.
3) We formulate the CQs for the uncovered needs.
4) We move to the next resource.
5) If the representation needs have been previously spotted in another resource, we omit them and move to a new representation need.
6) We formulate the CQs for the new uncovered needs.

Table 1. Characteristics of the analyzed terminological resources.

Name	Description	Domain	Languages
International Electrotechnical Vocabulary (Electropedia)	Terminological resource produced by an international standardization organization in electrotechnology	Electrotechnology	Multilingual (more than 15 languages)
TERMCAT Terminologia Oberta Platform	Terminological glossaries created by terminology experts in Catalonia, Spain, with standardization purposes	Multidisciplinary	Multilingual (4 languages)

[17] https://termcoord.eu/terminology-websites/

EcoLexicon	multimodal terminological knowledge base created by Terminology experts with conceptual, graphical and linguistic information	Environmental Sciences	Multilingual (6 languages)
InterActive Terminology for Europe (IATE)	Terminological database collaboratively created by translators and terminologists to support the drafting and translation process of EU legislation	Multidisciplinary	Multilingual (24 languages)

The International Electrotechnical Vocabulary.[18] Also known as Electropedia, it is a representative example of a terminological resource that follows the tenets of the *General Terminology Theory (GTT)*, which tries to make specialized knowledge universal through the standardization and normalization of terms (Wüster 1985). Electropedia is a controlled vocabulary that contains more than 22,000 source terms in English and French, with equivalents in more than 15 languages and it is structured as a concept schema, where terms are organized under the range of abstract classes. Although it does not contain explicit relations amongst concepts, we do find related concepts in the form of hyperlinks inside notes, definitions, and examples. Figure 1 shows an example of a term record related to the term *software*.[19] From this example, we have derived the Competency Questions presented in Table 2, in which we also include examples of properties and classes from existing Semantic Web models that provide means to represent each type of data.[20]

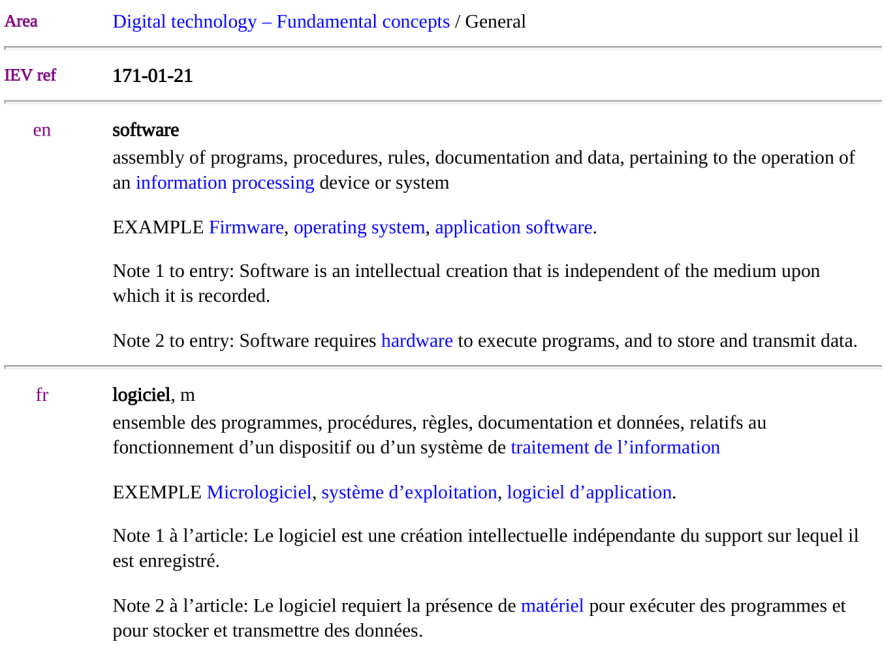

Figure 1. Example of the term entry *software* in Electropedia.

[18] https://www.electropedia.org/

[19] https://www.electropedia.org/iev/iev.nsf/display?openform&ievref=171-01-21

[20] These examples do not intend to be exhaustive but illustrative.

Table 2. CQs derived from the analysis of the data contained in Electropedia.

Competency Questions	Related Elements in Figure 1	Models	Classes/Properties
CQ1) What is the domain of the concept?	Area	DublinCore	dcterms:subject
CQ2) What is the identifier for the concept within the resource?	IEV ref	DublinCore	dcterms:identifier
CQ3) What are the terms that delineate the concept?	software (en), logiciel (fr)	OntoLex-lemon, SKOS	ontolex:LexicalEntry, skos:prefLabel, skos:alternativeLabel
CQ4) What is the language of a term?	en, fr	DublinCore	dcterms:language
CQ5) What is the definition of a term?	en: assembly of programs […] fr: ensemble des programmes […]	SKOS	skos:definition
CQ6) What are examples of a term?	EXAMPLE	SKOS, RDF Schema	skos:example, rdf:seeAlso
CQ7) What are the notes related to the term?	Note 1, Note 2	SKOS	skos:note
CQ8) What are the hyperlinks contained in the entry?	information processing, firmware, operating system […]	DublinCore	dcterms:identifier

TERMCAT Terminologia Oberta.[21] This platform is developed and maintained by the Centre for Catalan Terminology, in charge of the promotion, elaboration and dissemination of terminological resources, as well as the normalization of Catalan neologisms. The set of terminological resources accessible from the TERMCAT portal are multilingual in Catalan and Spanish, and usually contain equivalents in English and French.

This resource is closely related with the *Communicative Terminology Theory*, which gives special importance to the context in which the term is used (Cabré 2003; Montero-Martínez and Faber-Benítez 2009). For this reason, term entries in this resource tend to contain notes, as shown in Figure 2.[22] In this case, the note is not only a string of text, but may contain additional and relevant information, such as links to the sources from which the information was extracted and hyperlinks to related concepts, for instance. From this example, we raise the CQs presented in Table 3.

[21] https://www.termcat.cat/es/terminologia-oberta

[22] https://www.termcat.cat/es/diccionaris-en-linia/173/fitxa/Mjc3MDM4OA%3D%3D

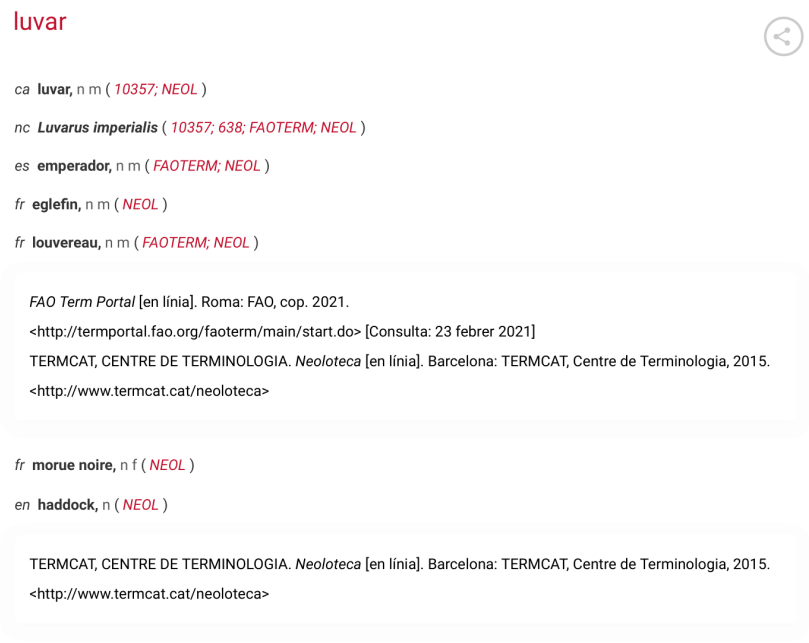

Figure 2. Example of the term entry *haddock* in the Terminologia Oberta platform from TERMCAT.

Table 3. CQs derived from the analysis of the data contained in TERMCAT.

Competency Questions	Related Elements in Figure 2	Models	Classes/Properties
CQ9) What are the equivalents of a term?	ca: luvar nc: Luvarus imperialis es: emperador […]	Vartrans	vartrans:Translation
CQ10) What is the part of speech of a term?	n (noun)	Lexinfo	lexinfo:PartOfSpeech
CQ11) What is the gender of a term?	ca: m (masculine) fr: f (feminine)	Lexinfo	lexinfo:gender
CQ12) What additional information is shown in the note of a term?	http://termportal.fao.org /faoterm/main/start.do http://termcat.cat/neoloteca	SKOS	skos:note dc:identifier
CQ13) What is the source of the term?	ca: luvar (10357; NEOL) fr: Luvarus imperialis (10357; 638; FAOTERM; NEOL)	DublinCore	dcterms:source

EcoLexicon.[23] This recognized terminological resource (Faber, León-Araúz and Reimerink 2016) is known to be the practical application of the *Frame-Based Terminology Theory (FBT)*, which exposes the idea that *frames* need to be created for a complete understanding of a term within a domain (Faber-Benítez, Márquez-Linares and Vega-Expósito 2005). EcoLexicon is a multilingual resource in Environmental Science with approximately 5000 concepts. One of its key features is the identification of hierarchical and non-hierarchical relations amongst its concepts, such as in Figure 3, in

[23] https://ecolexicon.ugr.es

which we observe the relations part-of, made-of, type-of and located-at. Also, on some occasions, concepts may have resources attached, which can be of different natures: images, videos, web pages and documents (as shown in Figure 4). This type of multimodal information could be represented with the Multimodality Module of OntoLex, that is now currently being discussed by the Ontology Lexica Community members.[24] From these examples, we identified the CQs presented in Table 4.

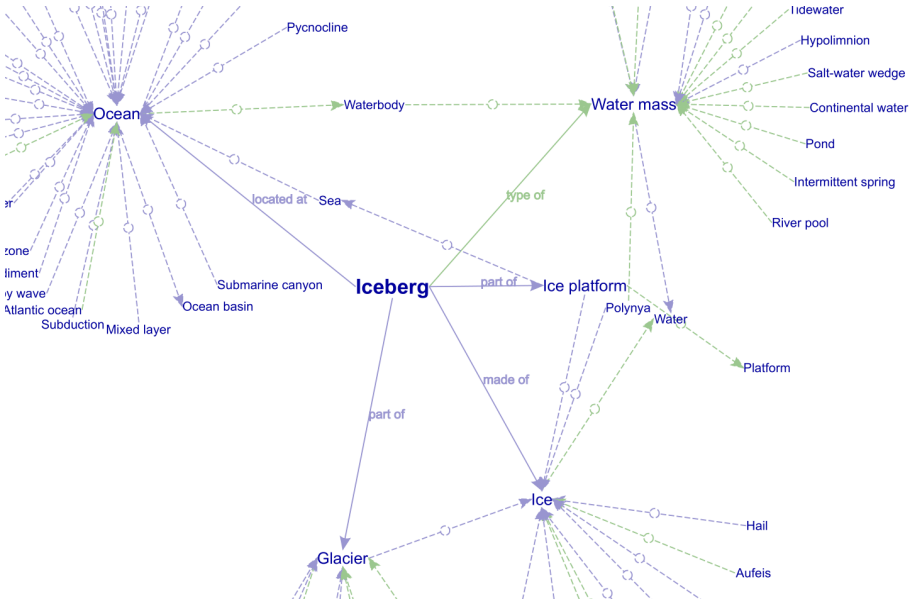

Figure 3. Example of the relations for the term *iceberg* in Ecolexicon.

Resource information

Title:	Iceberg
Description:	Icebergs are enormous. They can be taller than the Statue of Liberty and as big as a small state. Normally, only a tiny part of an iceberg can be seen from the surface of the water. Thousands of these massive glacial chunks are formed each year and many are carried hundreds of miles by ocean currents until they gradually melt away.
Image:	
Source:	http://express.howstuffworks.com/gif/wq-iceberg-underwater.jpg

Figure 4. Example of the resource information of the term *iceberg* in EcoLexicon.

Table 4. CQs derived from the analysis of the data contained in EcoLexicon.

Competency Questions	Related Elements in Figures 3 and 4	Models	Classes/Properties
CQ14) What is the image that explains the concept?	Image	Foaf, BIBO	foaf:Image, bibo:Image, foaf:img

[24] https://www.w3.org/community/ontolex/wiki/Multimodality

CQ15) What is the source of the image?	Source	DublinCore	dcterms:source
CQ16) What are the conceptual relations amongst concepts?	part of, made of, type of […]	OWL, RDF, SKOS, DBpedia OWL	owl:partOf, rdf:type, skos:narrower, skos:broader, dbpedia-owl:locatedInArea

Interactive Terminology for Europe.[25] The rest of our Competency Questions are formulated based on the examination of the most representative terminological database in Europe, IATE. This resource has recently evolved to IATE2, an improved version with modernized technologies, more interoperability, and standard data structures, which contains almost one million entries (each representing concepts) and around eight million terms (representing the different designations for the same concept) (Zorrilla and Fontenelle 2019).

For instance, when searching for the term *contract notice* in IATE, we get several term variants with different reliability degrees that, as in the term entry shown in Figure 5, are represented with stars (from 1 to 4).[26] For instance, *contract notice* has four stars while *tender notice* has only two, which indicates that the first term is more reliable than the latter. Moreover, IATE also provides evaluations of terms to indicate if a term is preferred, admitted, deprecated, obsolete or proposed.[27] In Figure 5, we see that the term *anuncio de contrato* is evaluated as *admitted*, while the term *tender notice* is evaluated as *obsolete*, which is in line with its low reliability degree. Apart from CQs that have already been formulated, namely domain, languages, and translation equivalents, we also propose those in Table 5.

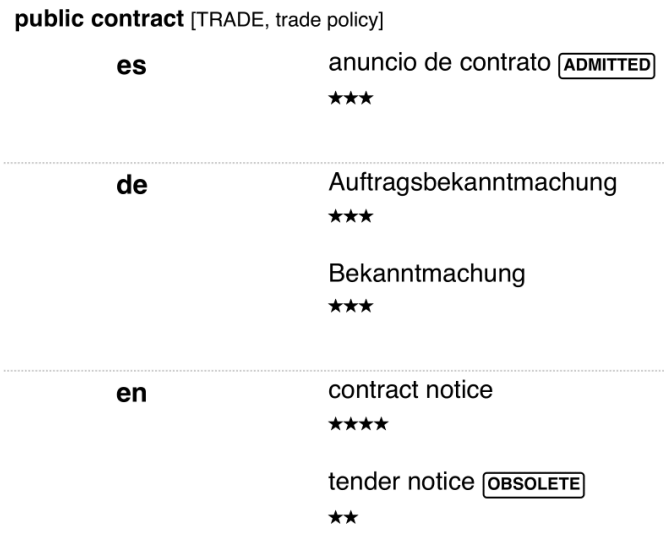

Figure 5. Example of the reliability and evaluation information of the term *contract notice* in IATE.

[25] https://iate.europa.eu/

[26] https://iate.europa.eu/entry/result/754472/en-de-es

[27] https://iate.europa.eu/fields-explained

Table 5. CQs derived from the analysis of the data contained in IATE (1).

Competency Questions	Related Elements in Figure 5	Models	Classes/Properties
CQ17) What is the evaluation of a term?	es: anuncio de contrato, ADMITTED en: tender notice, OBSOLETE	Lexinfo	lexinfo:normativeAuthorization
CQ18) What is the reliability of a term?	en: contract notice, four stars: tender notice, two stars	TBX2RDF	tbx:reliabilityCode
CQ19) What are the variants of a term?	de: Auftragsbekanntmachung, Bekanntmachung en: contract notice, tender notice	SKOS, Lexinfo, vartrans	skos:altLabel, lexinfo:synonym, vartrans:LexicalRelation, vartrans:SenseRelation

Additionally, in the same term entry of this resource, we may also see notes at different levels, such as in Figure 6, where additional information about a definition is given. Other information that can be found throughout IATE entries is the term type and contextual information, as in Figure 7. Consequently, we extracted the CQs exposed in Table 6.

Definition:

individual notice in which contracting authorities or contracting entities call on economic operators to compete for a public contract and set out their needs and requirements and other information depending on the type of procurement procedure

Definition reference:
Council-EN, based on
- Directive 2014/24/EU on public procurement, 02014L0024-20160101/EN

- Directive 2014/25/EU on procurement by entities operating in the water, energy, transport and postal services sectors CELEX:02014L0025-20160101/EN

Note:
Depending on the type of procurement procedure, other notices which can be used as a means of calling for competition are:
- prior information notice

- periodic indicative notice

- notice on the existence of a qualification system

Figure 6. Example of a definition note of the term *contract notice* in IATE.

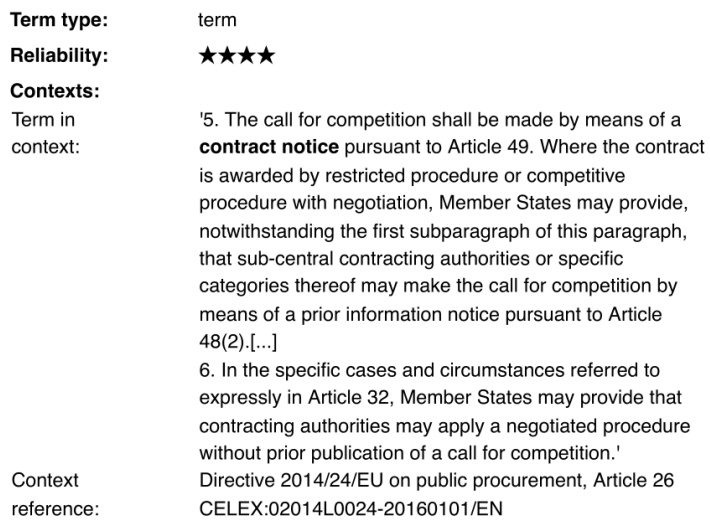

Figure 7. Example of additional information (reliability, context, and context source) of the term *contract notice* in IATE.

Table 6. CQs derived from the analysis of the data contained in IATE (2).

Competency Questions	Related Elements in Figures 6 and 7	Models	Classes/Properties
CQ20) What is the source of the definition of a term?	Definition reference	DublinCore	dcterms:source
CQ21) What is the note of the definition of a term?	Note	SKOS	skos:note
CQ22) What is the term type?	Term Type	Lexinfo	lexinfo:TermType
CQ23) What is the usage context of a term?	Context. Term in context.	Lexicog	lexicog:UsageExample
CQ24) What is the source of the usage context of a term?	Context reference	DublinCore	dcterms:source
CQ25) What is the language level note (usage) of a term?	Note[28]	OntoLex	ontolex:usage

Finally, we realized it is not coherent to propose a model for terminologies whose central class is a *LexicalConcept*, since this class is aimed at lexicographical resources. We consider necessary the addition of the *TerminologicalConcept* figure to the model, as a sister class of *LexicalConcept*, to maintain consistency. Both, as subclasses of skos:Concept, inherit the same attributes, although certain Termlex properties can only be attached to the *TerminologicalConcept* class. Therefore, we raise the last question: **CQ26) What is the Terminological Concept of a term?**

[28] Notes at this level can refer to the definition or to the usage of a term.

4. Proposal: Termlex

Analyzing the existing models to represent Linguistic Linked Data, we observe that SKOS is the most adopted vocabulary to represent and structure thesauri in RDF, whereas TBX remains the main format in terminology management tools. However, we perceive a trend towards terminological resources that combine different types of information: conceptual, terminological, lexicographical, multimodal, etc., and believe that a more comprehensive model might be required to cover all representational needs. For this reason, we deem it adequate to use a more comprehensive vocabulary such as OntoLex-lemon. Moreover, in the latest version (see Figure 8), OntoLex introduces a link to SKOS in the form of a subclass of skos:Concept, labelled ontolex:LexicalConcept, that serves as a bridge between the representation of lexical data and conceptual information.

Therefore, the main purpose of this proposal is to complement the OntoLex-lemon vocabulary with the necessary classes and properties to achieve a twofold purpose: to represent the information usually contained in traditional terminological resources and thesauri (IATE, TermCoord glossaries, EuroVoc, etc.), and to model terminological resources that might be created in a semi-automatic fashion by leveraging data available in resources in the LLOD environment, thus exposed as Linked Data.

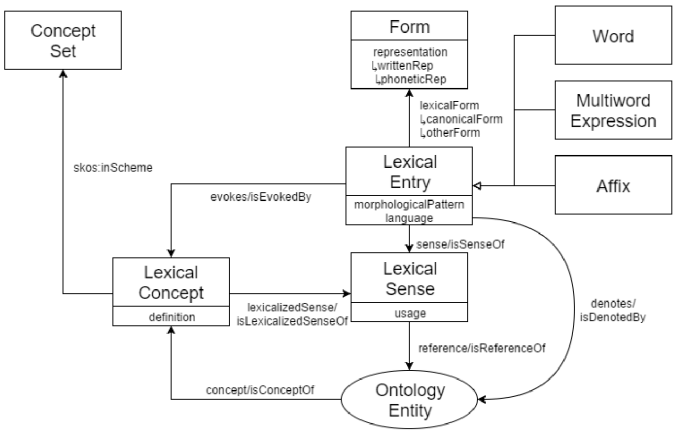

Figure 8. Current core diagram of OntoLex-lemon.

After the analysis presented in Section 3 of the type of information contained in terminological resources, we have identified two groups of Competency Questions:

- **Group 1**: CQ1, CQ2, CQ3, CQ4, CQ6, CQ8, CQ9, CQ10, CQ11, CQ14, CQ16, CQ19, CQ22 and CQ23 can be represented by current modelling solutions and do not need to be addressed in this work.
- **Group 2**: CQ5, CQ7, CQ12, CQ13, CQ15, CQ17, CQ18, CQ20, CQ21, CQ24, CQ25, and CQ26 have not been or have been partially covered by previous models.

Therefore, in this paper, we approach Group 2 of Competency Questions, those that have not yet been fully addressed. More specifically, we encountered the following issues:

- Elements such as definitions or notes are much more than just a string of text, and current specifications (skos:definition and skos:note) do not provide means to capture the semantics represented by each type of note (**CQ5**, **CQ7**, **CQ12**, **CQ20**, **CQ21** and **CQ25**). Notes and definitions may contain pointers to other data, in the form, for instance, of links amongst term IDs, links to the sources of the information, or links to notes with additional information.
- Likewise, information about the source of the data needs to be provided, since the various data fields of the term record may have different sources. Current vocabularies only offer properties to model this (dc:source), meaning that only one string of text can be represented, and no additional properties such as the author or the creation date can be collected. Also, chained provenance information may need to be captured: we may want to declare that the definition of a given term comes from IATE, which, in turn, draws it from a European regulation (**CQ13**, **CQ15** and **CQ24**).
- The trustworthiness of terms is a key usage indicator and a standardized way to model it should be provided, which is not the case with current vocabularies; defining a universal standard to harmonize this indicator is a complex effort (**CQ17** and **CQ18**).
- Since OntoLex-lemon is originally intended to model lexicographic resources, there is a need to cover the representation requirements of terminological resources (**CQ26**).

Consequently, the objectives put forward to complement the current version of OntoLex-lemon are varied. First, we aim to accurately represent the origin (provenance, source, reference) of certain information items traditionally contained in terminologies. The rationale behind the inclusion of the references or sources of information from which the terms themselves, the definitions or the usage contexts have been obtained or harvested, is that it adds credibility to the information contained in the resource. As claimed in the recently updated IATE User's Handbook[29] regarding the *credibility of entries*:

> A well thought-out IATE entry tries to give users as much information as possible to allow them to judge whether the proposed solution is appropriate and credible. It must also allow other terminologists wishing to work on the entry to delimit the concept clearly, by providing references to the relevant sources consulted.

Additionally, users are encouraged to include *authoritative, credible sources* (see section 4.3.5 of the Handbook), since it demonstrates the reliability of the term in question, a parameter used later to assess the reliability of the information contained in the resource. Thus, the Termlex proposal also covers term reliability and term evaluation.

By proposing such complementary classes to the current version of the OntoLex-lemon vocabulary, we would also allow users to account for the origin or source of linguistic information items obtained from resources in the Linguistic Linked Open Data cloud, when creating terminological resources (or any other type of linguistic resources) in a semi-automatic mode and reusing available data in the cloud. In this scenario, terminological definitions and their sources gain relevance, and a single triple does not suffice to capture all key information.

[29] https://iate.europa.eu/assets/IATE_Handbook_public.pdf

Considering that neither the specification of skos:definition nor the specification of dc:source restricts the object to be raw text, we recommend definitions and sources to be defined as resources (classes), with further attributed properties. With the purpose of harmonizing the properties to be used for definition and source, we believe a few classes and properties should be specified, thus favoring the interoperability of OntoLex-lemon implementations.

This proposal follows the path of previous modules to complement or extend OntoLex-lemon that have been discussed and implemented after the release of the model final specification in May 2016. Examples of such modules are: the Lexicog module[30] dedicated to the modelling of lexicographic data, the Morphology module[31] aimed at modelling the formation and decomposition of lexical entries and forms and the FrAC module,[32] to represent frequency, attestation and corpus information.

4.1. Termlex Development Methodology

When developing an ontology (model or vocabulary), it is highly recommended to follow an established ontology development methodology. In this proposal, two methodologies have been adopted:

- The NeOn methodology (Suárez-Figueroa, Gómez-Pérez and Fernández-López 2012), which suggests different scenarios for ontology engineering.
- The LOT (Linked Open Terms) methodology (Poveda-Villalón, Fernández-Izquierdo et al. 2022), which is an agile methodology, mostly industry-oriented and in line with the NeOn principles.

Both methodologies encourage developing an Ontology Requirement Specification Document (ORSD),[33] to identify the purpose, scope, implementation language, intended end-users, intended uses and ontology requirements. We have developed an ORSD for the Termlex model and it is shown in Table 7, containing the Competency Questions described in Section 3, amongst other relevant data about the model.

The ORSD is essential for the subsequent *ontology implementation* step, which in this case was done using the Protégé application. Once the ontology is implemented, it is evaluated using OOPS! (OntOlogy Pitfall Scanner!), a tool to identify the most common mistakes when developing ontologies (Poveda-Villalón et al. 2014). OOPS! is published as a web service[34] and is based on a catalogue of 41 pitfalls[35] derived from the analysis of almost 700 ontologies. The results of the evaluation are publicly available in the Termlex model's website.[36]

The next step, according to the LOT methodology (see Figure 9), is the *ontology publication*, which implies the generation of documentation, that in this case has been

[30] https://www.w3.org/community/ontolex/wiki/Lexicography
[31] https://www.w3.org/community/ontolex/wiki/Morphology
[32] https://www.w3.org/community/ontolex/wiki/Frequency,_Attestation_and_Corpus_Information
[33] https://github.com/oeg-upm/ORSD-template
[34] https://oops.linkeddata.es/
[35] https://oops.linkeddata.es/catalogue.jsp
[36] https://termlex.oeg.fi.upm.es/

published using WIDOCO,[37] a tool that helps automatically publish customized documentation of a given ontology (Garijo 2017). The final step in the methodology is the *ontology maintenance*. Since this is still a proposal that needs to be discussed and made official, this step is regarded as future work.

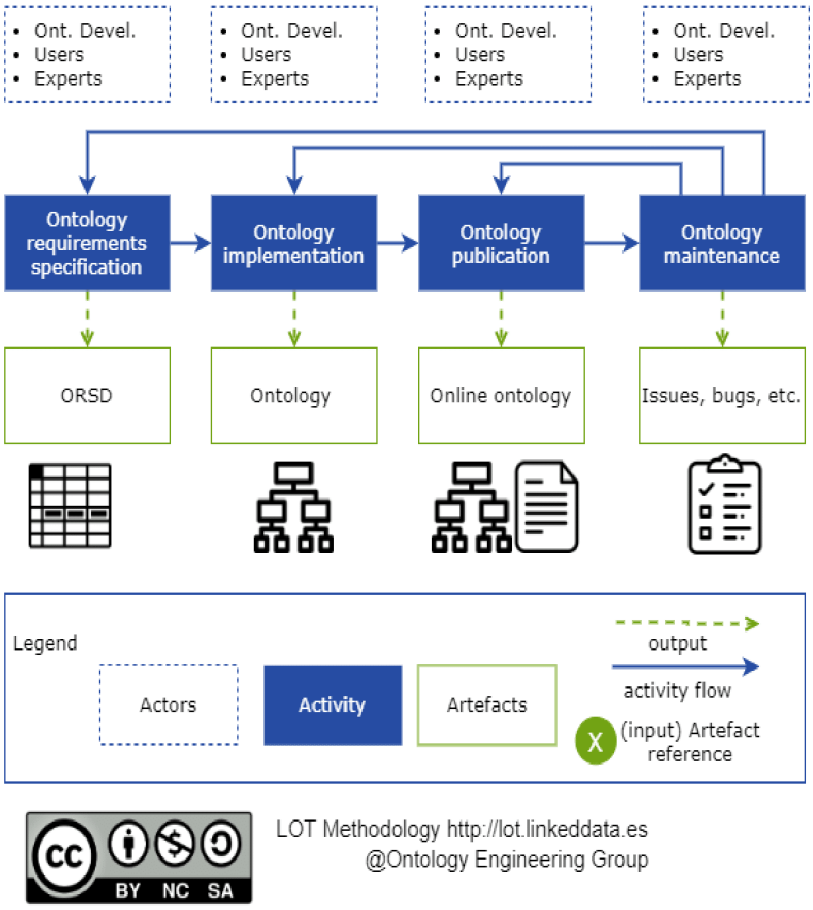

Figure 9. LOT methodology from Poveda-Villalón et al. (2022)

4.2. Termlex Core

The Termlex Core diagram (Figure 9) shows the classes and properties described in this proposal. Boxes represent classes and arrows represent properties. In this modelling approach each term is represented by a TerminologicalConcept, which is the main component of the model. In the left part of the figure, the OntoLex core diagram is shown (lines colored in blue). In the right part, the proposed Termlex elements are exposed (lines colored in green).

[37] https://github.com/dgarijo/Widoco

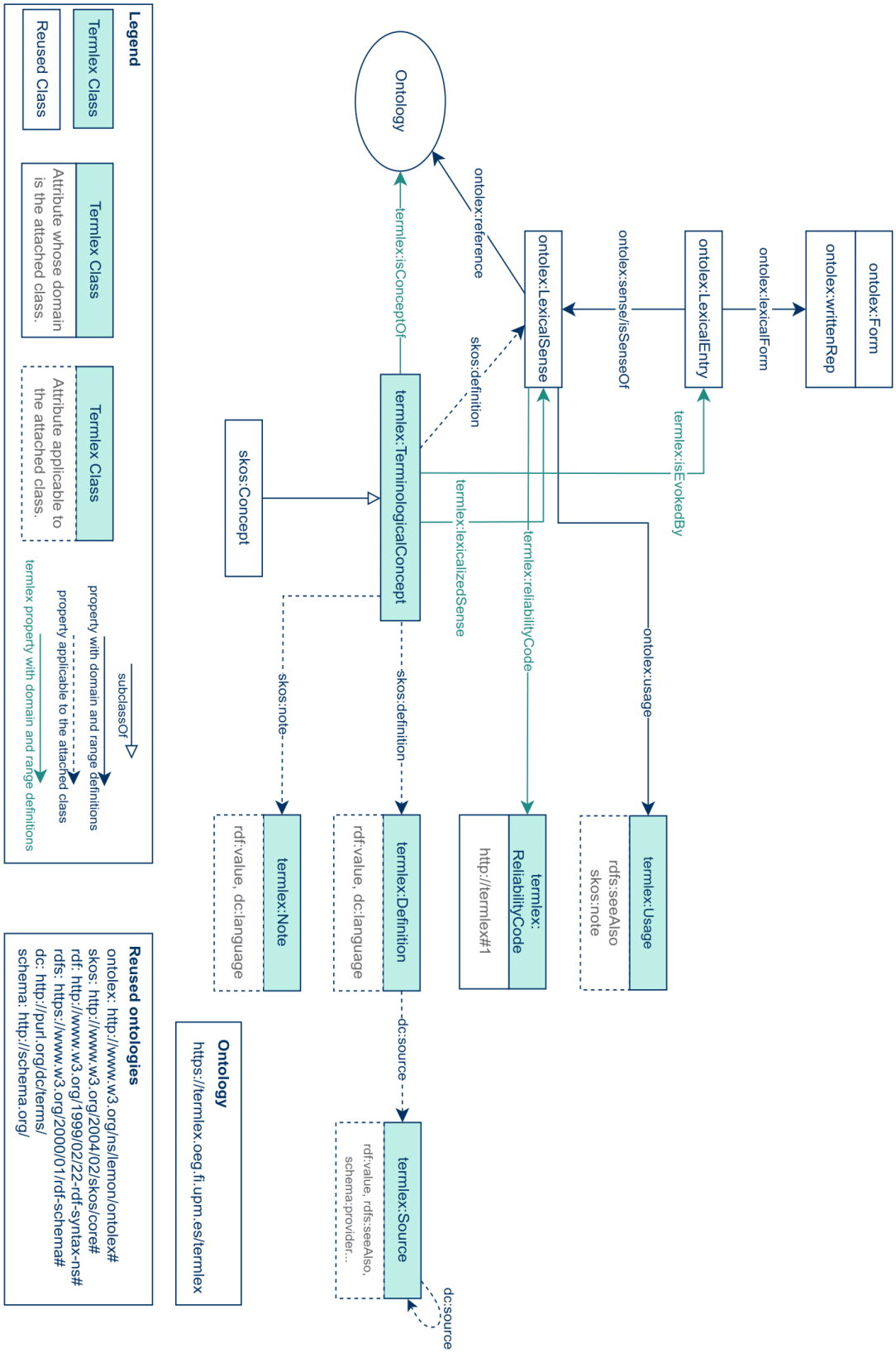

Figure 10. Termlex core diagram.

Table 7. Termlex Ontology Requirement Specification Document.

<table>
<tr><th colspan="2">Termlex Ontology Requirement Specification Document</th></tr>
<tr><th colspan="2">1. Purpose</th></tr>
<tr><td colspan="2">The purpose of the Termlex module is to complement the OntoLex-lemon vocabulary with the necessary classes and properties for a comprehensive and accurate publication of terminologies in RDF.</td></tr>
<tr><th colspan="2">2. Scope</th></tr>
<tr><td colspan="2">The scope of this module is limited to cover the gaps that OntoLex and its current extensions, may they be official or not (Vartrans, Lexicog, FrAC, Morphology, Multimodality, Valency and Semantics module), are not able to fulfil with regard to the representation of both traditional and modern terminologies, specifically: i) the representation of definitions and related information; ii) the representation of notes and related information; iii) the representation of sources and related information; iv) the representation of term reliability and term usage and v) the cohesion of the proposed extension with the OntoLex model.</td></tr>
<tr><th colspan="2">3. Implementation Language</th></tr>
<tr><td colspan="2">OWL</td></tr>
<tr><th colspan="2">4. Intended End-Users</th></tr>
<tr><td colspan="2">User 1. Linguists and terminologists. Preferably with technical skills.
User 2. Computer Scientists.
User 3. Other users with terminological and semantic web needs.</td></tr>
<tr><th colspan="2">5. Intended Uses</th></tr>
<tr><td colspan="2">Use 1. Conversion of traditional terminological resources into RDF maintaining all the original pieces of information.
Use 2. Representation of terminologies generated through semi-automatic processes from heterogeneous language resources.</td></tr>
<tr><th colspan="2">6. Ontology Requirements</th></tr>
<tr><th colspan="2">a. Non-Functional Requirements</th></tr>
<tr><td colspan="2">NFR 1. The ontology (Termlex module) needs to be used in combination with SKOS, the OntoLex vocabulary and its related extensions.
NFR 2. The ontology shall be published online with standard documentation.
NFR3. This ontology is a profile of the OntoLex-lemon vocabulary.[38]</td></tr>
<tr><th colspan="2">b. Functional Requirements: Competency Questions</th></tr>
<tr><td>CQ5) What is the definition of a term?
CQ7) What are the notes related to a term?
CQ12) What additional information is shown in the note of a term?
CQ13) What is the source of a term entry?
CQ15) What is the source of the image?
CQ17) What is the evaluation of a term?</td><td>CQ18) What is the reliability of a term?
CQ20) What is the source of the definition of a term?
CQ21) What is the note of the definition of a term?
CQ24) What is the source of the usage context of a term?
CQ25) What is the language level note (usage) of a term?
CQ26) What is the Terminological Concept of a term?</td></tr>
</table>

Note that in this proposal we follow the recommendation of the methodologies to reuse, whenever possible, the existing models to represent linguistic data: Lexinfo, OntoLex-lemon, Lexicog Module, amongst others. This proposal is intended to cover existing gaps and to avoid redundancy at the same time and does not intend to be restrictive but complementary. A detailed description of the classes is included in Section 4.2.1, and properties have been described and exemplified in Section 4.2.2.

[38] https://www.w3.org/TR/dx-prof/

4.2.1. Classes

termlex:Definition A *definition* is understood as a sentence or clause that explains the meaning of a concept. As mentioned earlier, we propose to reify the property definition into a class, so that further statements can be made about it. In this manner, its authorship, validity, and provenance, for instance, can be asserted. The creation of the termlex:Definition class helps answering *CQ5) What is the definition of a term?*, *CQ20) What is the source of the definition of a term?* and *CQ21) What is the note of the definition of a term?*

Class URI: http://www.w3.org/ns/lemon/termlex#Definition

termlex:Note Notes are key elements of traditional term records, providing additional information, such as usage recommendations, domain data and references; they are considered valuable pieces of knowledge for language professionals. We propose to reify the *note* properties of other models such as SKOS, OntoLex and Lexinfo (skos:note, ontolex:usage, lexinfo:note) into a class, so that we can also assert its provenance, or the additional and relevant data that notes may contain.

Throughout the analysis of terminological resources, we found that notes can appear at different levels: term notes, definition notes and context notes. Since the skos:note property does not have either domain nor range, we suggest reusing it to link the class Note to the classes ontolex:LexicalEntry or termlex:TerminologicalConcept, to represent term notes, depending on the representation approach (see Section 5); to the termlex:Definition, to represent definition notes and to the lexicog:UsageExample, to represent context notes. Thus, we do not need to create three subclasses for each type of note, as for the Source class, since we can easily infer their provenance. The creation of the termlex:Note class answers *CQ7) What are the notes related to the term?, CQ12) What additional information is shown in the note of a term? and CQ21) What is the note of the definition of a term?*

Class URI: http://www.w3.org/ns/lemon/termlex#Note

termlex:Source Like definitions, sources play a very important role in this modelling approach. Especially when terminologies are generated from multiple resources, as described in the ORSD, it is crucial to maintain the traceability of the different terminological data (may they be definitions, term notes, term contexts, etc.). With the automation of the terminology creation process, we may distinguish between two types of sources:

- **Intermediate Sources**: not direct sources but information providers, such as existing linguistic resources from which information is retrieved (IATE, for instance) or applications (a Definition Extractor).
- **Original Sources**: direct sources, meaning corpora (i.e. European Legislation), organizations (i.e. European Commission) or individuals (i.e. John Doe, European terminologist)

Class URI: http://www.w3.org/ns/lemon/termlex#Source

In this case, however, we think it is not necessary to make this distinction explicit by creating two subclasses as in the previous case, since the difference between intermediate and original sources can easily be deduced from the representation of chained sources. The creation of the class termlex:Source answers *CQ13) What is the source of a term entry?, CQ20) What is the source of the definition of a term? and CQ24) What is the source of the usage context of a term?*

termlex:ReliabilityCode Previous work on the representation of terminologies as Linked Data (Cimiano et al. 2015) proposed an ontology based on the TBX specification which used the property tbx:reliabilityCode to represent this kind confidence rating that terminologists assign to terms.[39] However, the domain is ontolex:LexicalEntry, and the property admits any type of rating. Following the guidelines of IATE,[40] we propose a ReliabilityCode class that answers *CQ18) What is the reliability of a term?*. This class points at a fixed set of numerical values, 1-4, such as:

http://www.w3.org/ns/lemon/termlex#ReliabilityCode#1

termlex:Usage Apart from the three levels of notes mentioned before, throughout IATE entries we can also find *language level notes* that may be attached to the same Terminological Concept, but containing different information. From our analysis, and based on the IATE User's Handbook, we realized that the kind of information present in this section is directly related to the usage of the term. We propose the reification of the property offered by Ontolex to the class termlex:Usage, since in IATE we find additional data such as links to other resources and identifiers. This class answers *CQ25) What is the language level note (usage) of a term?*

Class URI: http://www.w3.org/ns/lemon/termlex#Usage

termlex:TerminologicalConcept As explained at the end of Section 3, an important contribution of this proposal is the addition of termlex:TerminologicalConcept as a sister class of ontolex:LexicalConcept, to maintain the consistency throughout the model. The addition of this class does not affect the OntoLex vocabulary: we are not redefining any property related to this class but adding new ones to account for their domains and ranges, as described in the following section. This class answers *CQ26) What is the Terminological Concept of a term?*

4.2.2. Properties

termlex:isEvokedBy We suggest using this property to indicate that a TerminologicalConcept can be designated by one LexicalEntry. It is, therefore, a sister property of ontolex:isEvokedBy, with a different domain.

Property URI: http://www.w3.org/ns/lemon/termlex#isEvokedBy
Domain: termlex:TerminologicalConcept
Range: ontolex:LexicalEntry
Inverse property: termlex:evokes

---

[39] https://github.com/cimiano/tbx2rdf/blob/master/ontology/tbx.owl
[40] https://iate.europa.eu/fields-explained

termlex:lexicalizedSense Similarly, termlex:lexicalizedSense is a sister property of ontolex:lexicalizedSense, again, with a different domain. Since the central class of this model is the termlex:TerminologicalConcept, this property is required to attach senses to it.

Property URI: http://www.w3.org/ns/lemon/termlex#lexicalizedSense
Domain: termlex:TerminologicalConcept
Range: ontolex:LexicalSense
Inverse property: termlex:isLexicalizedSenseof

termlex:reliabilityCode Since we proposed the class termlex:ReliabilityCode with a fixed range of values, we also need a property to link this new class with the termlex:TerminologicalConcept to which it refers.

Property URI: http://www.w3.org/ns/lemon/termlex#reliabilityCode
Domain: termlex:TerminologicalConcept
Range: termlex:ReliabilityCode

5. Use Cases

Following the LOT methodology, we demonstrate the potential of the proposal through three use cases in which the Termlex model can be applied to satisfy the varied representation requirements. We therefore contemplate three different representation approaches:

1. **Simple resources**: In this category (Section 5.1), we include terminological assets with a narrow scope in terms of domain-specificity and types of data. Traditionally, this kind of resource has been manually elaborated and afterwards digitalized.
2. **Complex resources**: In this category (Section 5.2) we include bigger resources, with a wider scope in terms of domain-specificity and types of data. These resources include collaboratively generated terminologies and platforms that collect data from different resources.
3. **Automatically generated resources**: In this category (Section 5.3), we include resources that have been automatically generated with different NLP techniques, in which there is little or no human interaction.

Throughout the following sections, we point out the most important features that Termlex helps represent. Please note that we have not included every piece of data within the examples to avoid too complex diagrams that could lead to information overload.

5.1. Representing simple resources with Termlex

To exemplify this first approach, we take the term entry *software* from Electropedia that was already analyzed to extract the Competency Questions. The most important particularities of this example are:

- Two terms to describe the same concept, one in English and one in French.
- The same definition for the two terms, in both languages.
- Two notes for the concept, in both languages.

– Cross-references inside the definition and the note.

Figure 11 shows a simplified graphical representation of this first approach, which includes the following features:

– There is one termlex:TerminologicalConcept per term record, therefore, the same concept for both terms.
– We skip the class ontolex:LexicalSense, since there is no information that needs to be represented at this level.
– We represent the terms with the ontolex:LexicalEntry class, that is linked to the concept with the termlex:isEvokedBy property.
– Since there is a unified definition for the different terms, we represent it at the concept level, with the class termlex:Definition, which has two values in the different languages.
– The same is approach is used to represent the notes, with the class termlex:Note.

5.2. Representing complex resources with Termlex

In this second approach, we represent the term *train path* from IATE. We selected this entry due to its complexity: it contains every piece of data that can be found in an IATE entry. Thus, the particularities of this example are:

– As in the previous approach, we also have one concept per terminological entry. Therefore, several terms can describe one single concept.
– Similarly, there is one definition per concept, translated into the languages of the resource.
– The different terms, whether synonyms, term variants or translation equivalents, can have different sources and multiple notes.
– The definition may also have different sources and notes.
– In this resource, we also find information regarding the usage of the term, its reliability and its status.

Figure 12 shows a simplified graphical representation of this second approach, which includes the following features:

– There is one termlex:TerminologicalConcept per term record, therefore, the same concept for all the terms.
– In this case, we do need the ontolex:LexicalSense class, since there is information that needs to be represented at this level, specifically, information about usage, term reliability and status.
– The equivalent relations are also represented at the ontolex:LexicalSense level, by using the Vartrans module of OntoLex.
– Since there is a unified definition for the different terms, we represent it at the concept level, with the class termlex:Definition, that has two values in the different languages. However, the definition for this concept has several sources that are represented with three different termlex:Source classes.
– The same is approach is used to represent the notes, with the class termlex:Note.

5.3. Representing automatically generated resources with Termlex

This final approach reproduces a use case in which the information for a given term, again *train path*, has been automatically extracted from different resources, namely IATE, BabelNet[41] and Wikidata.[42] The particularities of this example are:

- Since the terms are automatically retrieved, we do not know if they refer to the same concept, therefore, we need to create one concept per term.
- These concepts are retrieved from different resources.
- Depending on the type of information retrieved, the modelling approach may change. In this case, we do not need to represent information at the sense level.
- The definitions may have been extracted from different resources.

Figure 13 shows a simplified graphical representation of this third approach, which includes the following features:

- We create a termlex:TerminologicalConcept per term.
- We do not use the ontolex:LexicalSense class.
- Therefore, different classes of termlex:Definition can be attached to each termlex:TerminologicalConcept.
- Each termlex:Definition has a termlex:Source class attached to it.
- To represent that a term is an equivalent of the other, we add a vartrans:translatableAs relations amongst the classes ontolex:LexicalEntry of each term.

6. Conclusions

In this paper, we tackle the representation of terminological data in Semantic Web formats. We start by analyzing a selection of authoritative and varied terminological resources (Electropedia, Termcat, EcoLexicon and IATE), with the goal of identifying their representation requirements. We exhaustively review state-of-the-art models to represent Linguistic Linked Data, aiming at establishing their scope and representation capabilities.

In this review, we identified several representation requirements related to different types of data present in term records. To cover those requirements, we propose a model based on the OntoLex-lemon vocabulary, which is currently the most comprehensive and applied model to represent linguistic data on the Web. This proposal is named Termlex and pursues standardizing the representation of both linguistic and conceptual information from terminologies in Semantic Web formats since there are currently no updated vocabularies that handle this.

The advantages of the publication of resources in Semantic Web formats and following the so-called Linked Data principles are numerous and varied. Firstly, these are structured formats, so each piece of information is categorized and can be easily retrieved, reused, and integrated in other resources. Secondly, since the core idea of these formats is to make information openly accessible, it can be easily updated by users, which can be translated into a lower risk of retrieving incorrect and deprecated information. Thirdly, thanks to the

[41] https://babelnet.org/

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

Linked Data paradigm, resources contain links to other resources, which allows users to quickly navigate from one resource to another, enhancing their reusability. Finally, due to the wide variety of available ontologies, almost everything can be represented, including the metadata of each entry in the resource, which is a valuable element to assess the quality and keep track of the information presented.

We exemplify the potential of Termlex through a series of use cases that propose three different modelling approaches depending on the type and complexity of the terminological data: 1) simple resources, 2) complex resources, 3) automatically generated resources. With these three use cases we intend to provide a preliminary set of guidelines and good practices that helps users implement Termlex on their data.

In conclusion, the purpose of Termlex is to complement the existing vocabularies, and the modelling approaches suggested are thought to help users understand the possibilities of this model; under no circumstances does this paper intend to be restrictive. Also, this model is not intended to cover every type of resource, since each resource has different types of data. We propose the classes and properties that allow the modelling of the data, the modelling decisions depend on the user. In line with previous W3C efforts, Termlex is a minimum viable proposal to homogenize this issue, and it should be extended for particular applications.

The Termlex model is available in an open GitHub repository, together with a series of examples in the form of diagrams and RDF Turtle files.[43]

Acknowledgements This work has been partially funded by the COST Action (European Cooperation in Science and Technology) through NexusLinguarum, the "European network for Web-centered linguistic data science" COST Action (CA18209) and by the project Knowledge Spaces: Técnicas y herramientas para la gestión de grafos de conocimientos para dar soporte a espacios de datos (Grant PID2020-118274RB-I00, funded by MCIN/AEI/ 10.13039/501100011033).

References

Berners-Lee, Tim. 2006. *Design Issues.* Last accessed: October 11, 2022. https://www.w3.org/DesignIssues/LinkedData.html.

Berners-Lee, Tim, James Hendler, and Ora Lassila. 2001. "The Semantic Web." *Scientific American*, 248(5), 34-43.

Bizer, Christian, Tom Heath, and Tim Berners-Lee. 2011. "Linked Data: The Story So Far." In *Semantic Services, Interoperability and Web Applications: Emerging Concepts: Emerging Concepts*, by Amit Sheth, 205-227. Hershey: IGI Global.

Bosque-Gil, Julia, Elena Montiel-Ponsoda, Jorge Gracia, and Guadalupe Aguado-de-Cea. 2016. "Terminoteca RDF: a Gathering Point for Multilingual Terminologies in Spain." *International conference on Terminology and Knowledge Engineering.* 136-146.

Bosque-Gil, Julia, Jorge Gracia, Elena Montiel-Ponsoda, and Guadalupe Aguado-de-Cea. 2016. "Modelling Multilingual Lexicographic Resources for the Web of Data: The

[43] https://github.com/pmchozas/termlex

K Dictionaries Case." *GLOBALEX 2016 Lexicographic Resources for Human Language Technology Workshop Programme.*

Cabré, María Teresa. 1993. *La Terminología. Teoría, Metodología, Aplicaciones.* Barcelona: Empuries.

Cabré, Maria Teresa. 2003. "Theories of Terminology: Their Description, Prescription and Explanation." *Terminology. International Journal of Theoretical and Applied Issues in Specialized Communication*, 9(2), 163-199.

Caracciolo, Caterina, Armando Stellato, Ahsan Morshed, Gudrun Johannsen, Sachit Rajbhandari, Yves Jaques, and Johannes Keizer. 2013. "The AGROVOC Linked Dataset." *Semantic Web*. 4(3), 341-348.

Chandrasekaran, Balakrishnan, John Josephson, and Richard Benjamins. 1999. "What are Ontologies and Why Do We Need Them?" *IEEE Intelligent Systems and Their Applications*, 14(1), 20-26.

Cimiano, Philipp, John McCrae, Víctor Rodríguez-Doncel, Tatiana Gornostay, Asunción Gómez-Pérez, Benjamin Siemoneit, and Andis Lagzdins. 2015. "Linked Terminologies: Applying Linked Data Principles to Terminological Resources." *Electronic Lexicography in the 21st Century.* 504-517.

Di Buono, Maria Pia, Philipp Cimiano, Mohammad Fazleh Elahi, and Frank Grimm. 2020. "Terme-a-LLOD: Simplifying the Conversion and Hosting of Terminological Tesources as Linked Data." *Linked Data in Linguistics.* 28-35.

Díez-Alvite, Luisa, Beatriz Pérez-León, Mercedes Martínez-González, and Javier Vicente Blanco-Dámaso. 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.

Faber, Pamela, Pilar León-Araúz, and Arianne Reimerink. 2016. "EcoLexicon: New features and Challenges." *Globalex Workshop at Language Resources and Evaluation Conference.* 73-80.

Faber, Pamela, Carlos Márquez-Linares, and Miguel Vega-Expósito. 2005. "Framing Terminology: A Process-Oriented Approach." *Meta: Journal des Traducteurs,* 50(4).

Francopoulo, Gil. 2013. *LMF Lexical Markup Framework.* Hoboken, New Jersey: John Wiley and Sons.

Fuertes-Oliveira, Pedro A., and Sven Tarp. 2014. *Theory and Practice of Specialised Online Dictionaries. Lexicography versus Terminography.* Berlin, Boston: De Gruyter.

Garijo, Daniel. 2017. "Widoco: a wizard for documenting ontologies." *Proceedings of the International Semantic Web Conference*, 94-102.

Gracia, Jorge, Marta Villegas, and Asunción Gómez-Pérez. 2018. "The Apertium Bilingual Dictionaries on the Web of Data." *Semantic Web*, 9(2), 231-240.

Gruber, Thomas R. 1993. "A Translation Approach to Portable Ontology Specifications." *Knowledge Acquisition,* 5(2), 199-220.

Hovenga, Evelyn. 2022. "Guideline and knowledge management in a digital world." In *Roadmap to Successful Digital Health Ecosystems. A Global Perspective.*, ed. by Evelyn Hovenga and Heather Grain, 239-270. Elsevier.

Ide, Nancy, and Jean Véronis. 1995. *Text Encoding Initiative: Background and Contexts.* Springer Science & Business Media.

McCrae, John, Guadalupe Aguado-de-Cea, Paul Buitelaar, Philipp Cimiano, Thierry Declerck, Asunción Gómez-Pérez, Jorge Gracia, et al. 2012. "Interchanging Lexical Resources on the Semantic Web." *Language Resources and Evaluation* 701–719.

Melby, Alan. 2012. "Terminology in the Age of Multilingual Corpora." *The Journal of Specialised Translation*, 18, 7-29.

Melby, Alan. 2015. "TBX: A Terminology Exchange Format for the Translation and Localization Industry." *Handbook of Terminology* 393-424.

Miles, Alistair, Brian Matthews, Michael Wilson, and Dan Brickley. 2005. "SKOS core: Simple Knowledge Organisation for the Web." *International Conference on Dublin Core and Metadata Applications.* 3-10.

Montero-Martínez, Silvia, and Pamela Faber-Benítez. 2009. "Terminological Competence in Translation." *Terminology. International Journal of Theoretical and Applied Issues in Specialized Communication*, 15(1), 88-104.

Montiel-Ponsoda, Elena, Guadalupe Aguado-de-Cea, Asunción Gómez-Pérez, and Wim Peters. 2011. "Enriching Ontologies with Multilingual Information." *Natural Language Engineering*, 17(3), 283-309.

Pastor Sánchez, Juan Antonio. 2016. "Proposal to Represent the UNESCO Thesaurus for the Semantic Web Applying ISO-25964." *Brazilian Journal of Information Studies: Research Trends*, 10(1), 1-8.

Pérez-Hernández, María Chantal. 2002. "Explotación de los Córpora Textuales Informatizados para la Creación de Bases de Datos Terminológicas Basadas en el Conocimiento." *Estudios de Lingüística del Español*, 18.

Poveda-Villalón, María, Alba Fernández-Izquierdo, Mariano Fernández-López, and Raúl García-Castro. 2022. "LOT: An Industrial Oriented Ontology Engineering Framework." *Engineering Applications of Artificial Intelligence*, 111, 104755.

Poveda-Villalón, María, Asunción Gómez-Pérez, ed. by Mari Carmen Suárez-Figueroa. 2014. "Oops!(ontology pitfall scanner!): An on-line tool for ontology evaluation." *International Journal on Semantic Web and Information Systems (IJSWIS)* 7-34.

Reymonet, Axel, Jérôme Thomas, and Nathalie Aussenac-Gilles. 2007. "Modelling Ontological and Terminological Resources in OWL DL." *From Text to Knowledge: The Lexicon/Ontology Interface - Workshop at ISWC07.* Busan: CEUR.

Roche, Christophe. 2012. "Ontoterminology: How to unify terminology and ontology into a single paradigm." *LREC 12 - Eighth International Conference in Laguage Resources and Evaluation* 21-27.

Roche, Christophe, Luc Damas, and Julien Roche. 2014. "Multilingual Thesaurus: The Ontoterminology Approach." *CIDOC 2014 (International Committee for Documentation) Access and Understanding. Networking in the Digital Era.*

Sager, Juan. 1990. *Practical Course in Terminology Processing.* Amsterdam: John Benjamins Publishing.

Stefaniak, Karolina. 2017. "Terminology Work in the European Commission: Ensuring High-quality Translation in a Multilingual Environment." *Quality aspects in institutional translation*, 8, 109-121.

Suárez-Figueroa, Mari Carmen, Asunción Gómez-Pérez, and Boris Villazón-Terrazas. 2009. "How to Write and Use the Ontology Requirements Specification Document." *OTM Confederated International Conferences "On the Move to Meaningful Internet Systems".* Springer. 966-982.

Suárez-Figueroa, Mari Carmen, Asunción Gómez-Pérez, and Mariano Fernández-López. 2012. "The NeOn Methodology for Ontology Engineering." In *Ontology Engineering in a Networked World*, ed. by Mari Carmen Suárez-Figueroa, Asunción Gómez-Pérez and Aldo Gangemi, 9-34. Heidelberg: Springer.

Temmerman, Rita. 2000. *Towards New Ways of Terminology Description. The sociocognitive approach.* Amsterdam/Philadelphia: John Benjamins.

Temmerman, Rita, and Koen Kerremans. 2003. "Termontography: Ontology building and the sociocognitive approach to terminology description." *Proceedings of CIL17.*

Uschold, Mike, and Michael Gruninger. 1996. "Ontologies: Principles, Methods and Applications." *The Knowledge Engineering Review* 93-136, 11(2).

Vandenbussche, Pierre-Yves, A Ghislain, María Poveda-Villalón, and Bernard Vatant. 2017. "Linked Open Vocabularies (LOV): a Gateway to Reusable Semantic Vocabularies on the Web." *Semantic Web*, 8(3), 437-452.

Wüster, Eugen. 1985. *Einführung in die allgemeine Terminologielehre und terminologische Lexikographie.* Copenhagen: Fachsprachliches Zentrum Handelshochschule.

Zapilko, Benjamin, Johann Schaible, Philipp Mayr, and Brigitte Mathiak. 2013. "TheSoz: A SKOS Representation of the Thesaurus for the Social Sciences." *Semantic Web*, 4(3), 257-263.

Zorrilla, Paula, and Thierry Fontenelle. 2019. "IATE 2: Modernising the EU's IATE Terminological Database to Respond to the Challenges of Today's Translation World and Beyond." *Terminology. International Journal of Theoretical and Applied Issues in Specialized Communication*, 25(2)146-174.

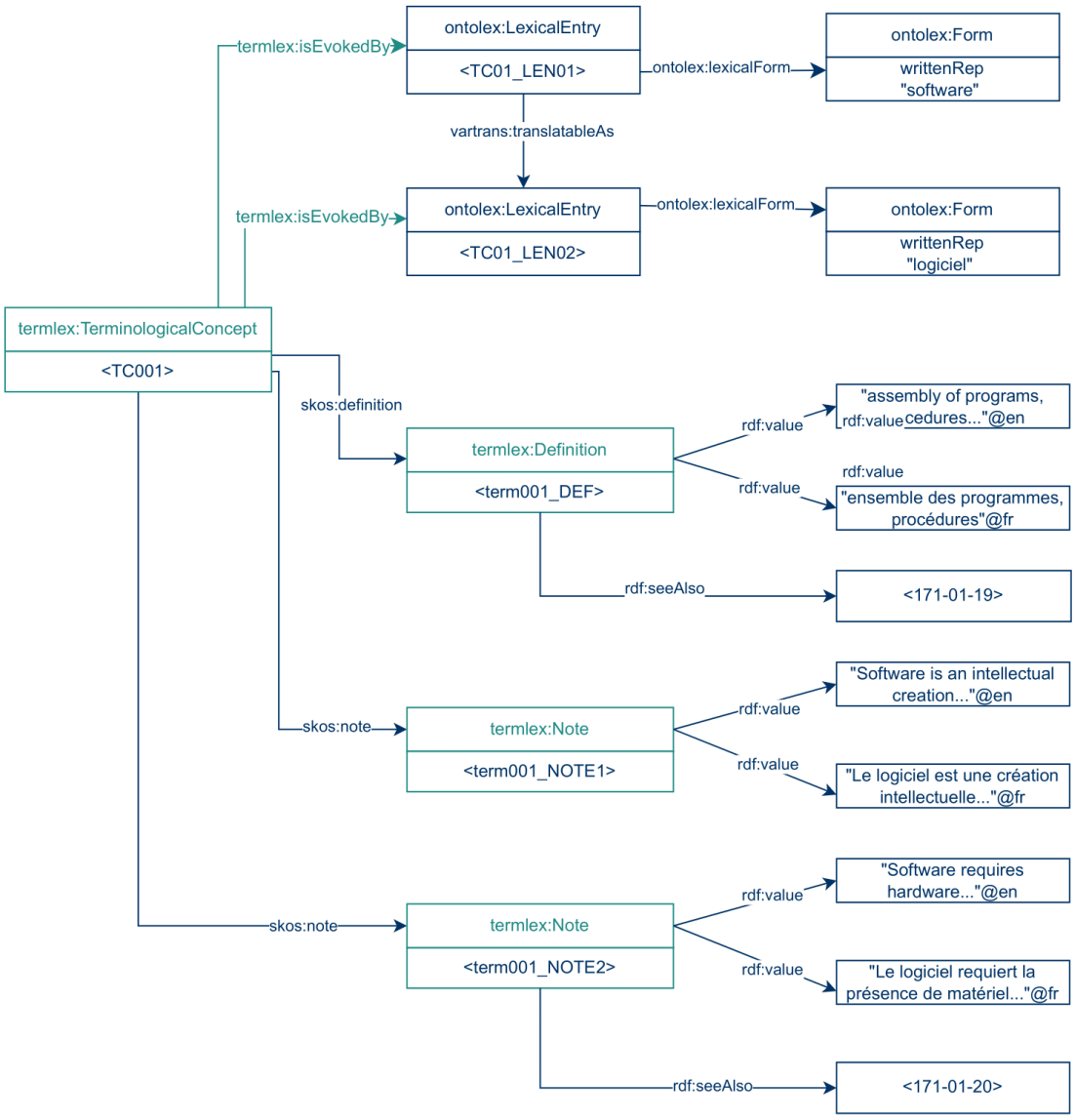

Figure 11. Simplified representation of the term entry *software* of Electropedia with Termlex (Approach 1).

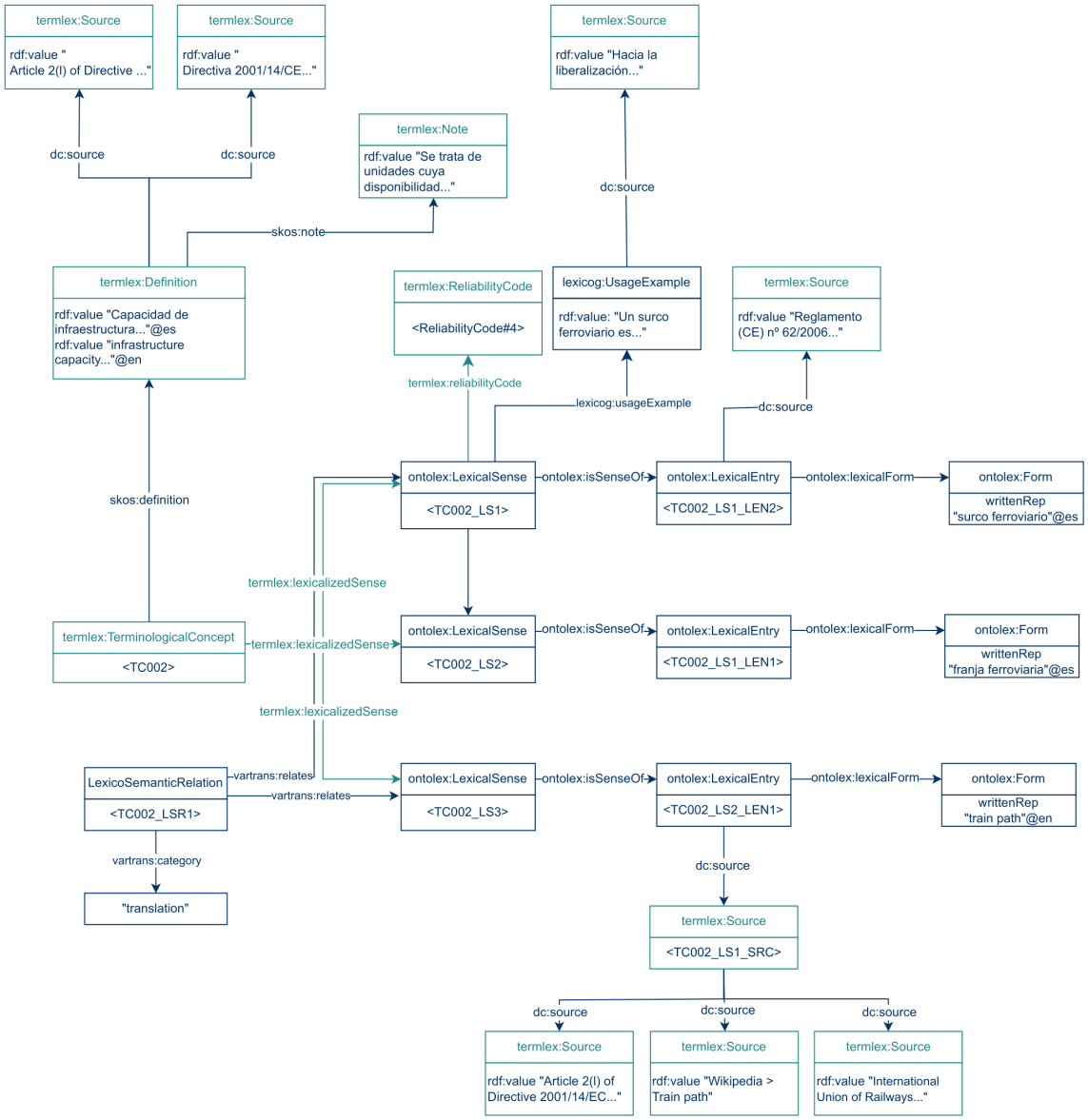

Figure 12. Simplified representation of the term entry *train path* of IATE with Termlex (Approach 2).

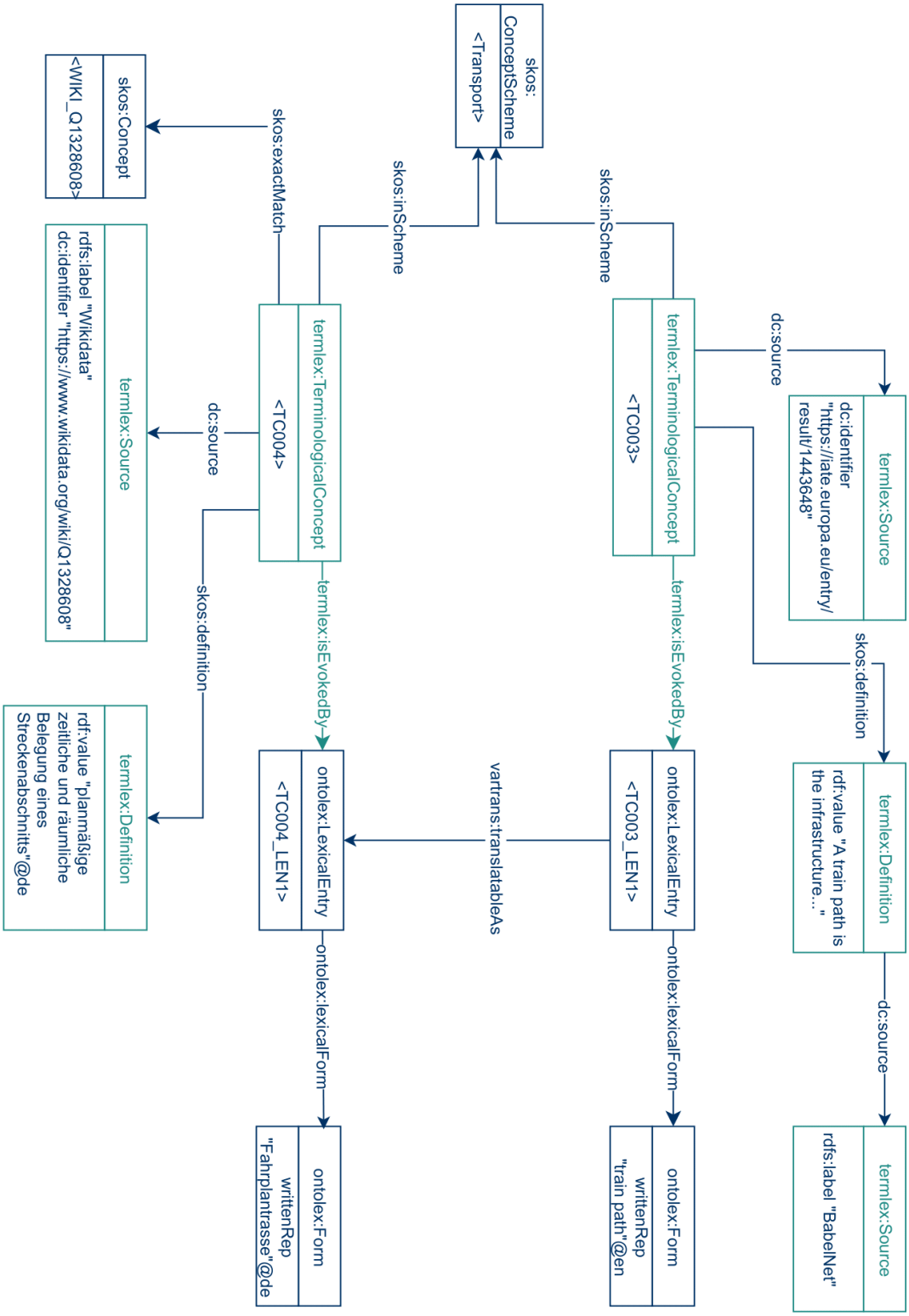

Figure 13. Simplified representation of an automatically generated entry for the term *train path* with Termlex (Approach 3).

Addresses for correspondence

Patricia Martín Chozas
Ontology Engineering Group, Escuela Técnica Superior de Ingenieros Informáticos,
Universidad Politécnica de Madrid
Campus de Montegancedo, s/n.
28660 Boadilla del Monte, Madrid
España
patricia.martin@upm.es

Thierry Declerck
Deutsches Forschungszentrum für Künstliche Intelligenz GmbH (DFKI)
Stuhlsatzenhausweg 3 Saarland Informatics Campus D 3_2
66123 Saarbrücken
Deutschland
declerck@dfki.de

Elena Montiel Ponsoda
Ontology Engineering Group, Escuela Técnica Superior de Ingenieros Informáticos,
Universidad Politécnica de Madrid
Campus de Montegancedo, s/n.
28660 Boadilla del Monte, Madrid
España
elena.montiel@upm.es

Víctor Rodríguez Doncel
Ontology Engineering Group, Escuela Técnica Superior de Ingenieros Informáticos,
Universidad Politécnica de Madrid
Campus de Montegancedo, s/n.
28660 Boadilla del Monte, Madrid
España
victor.rodriguez@upm.es