





LemonizeTBX: Design and Implementation of a New Converter from TBX to OntoLex-Lemon

Andrea Bellandi <andrea_dot_bellandi_at_ilc_dot_cnr_dot_it>, Institute for Computational Linguistics "A. Zampolli" CNR, Via Moruzzi 1, 56124, Pisa - Italy  <https://orcid.org/0000-0002-1900-5616>
Giorgio Maria Di Nunzio <giorgiomaria_dot_dinunzio_at_unipd_dot_it>, Department of Information Engineering, University of Padova, Via Gradenigo 6/b, 35131 Padova, Italy  <https://orcid.org/0000-0001-9709-6392>
Silvia Piccini <silvia_dot_piccini_at_ilc_dot_cnr_dot_it>, Institute for Computational Linguistics "A. Zampolli" CNR, Via Moruzzi 1, 56124, Pisa - Italy  <https://orcid.org/0000-0002-2584-0191>
Federica Vezzani <federica_dot_vezzani_at_unipd_dot_it>, Department of Linguistic and Literary Studies, University of Padova, Via Elisabetta Vendramini, 13 35137 Padova, Italy  <https://orcid.org/0000-0003-2240-6127>

Abstract

In this paper, we introduce LemonizeTBX, a converter that enhances interoperability between terminological and lexicographical frameworks, acknowledging their divergent data modelling approaches. First, we present the theoretical implications of a conversion from the TermBase eXchange (TBX) concept-oriented framework to the OntoLex-Lemon sense-centred standpoint within Semantic Web technologies. Then, we illustrate the prototype version of the converter, designed as an interactive tool engaging terminologists in the conversion process.

1. Introduction

In recent years, *Linked Data* (LD) has emerged as a highly promising approach to effectively integrate, interconnect, and semantically enrich different datasets, thus promoting interoperability and accessibility across the web [Frey and Hellmann 2021]. Terminologists therefore increasingly recognise the importance of publishing resources as LD to overcome siloed information management, and facilitate the reuse of terminological datasets in accordance with the FAIR principles [Wilkinson et al 2016]. In response to this paradigm shift, alongside the well-established ISO standard 30042:2019 about the *TermBase eXchange implementation format* (TBX),^[1] the OntoLex-Lemon data model has gained increasing importance within the terminology community, being widely recognized as the de facto standard for constructing RDF lexical resources (inter al. see [Bosque-Gil et al 2015]; [Rodriguez-Doncel et al 2018]; [Vellutino et al 2016]).^[2] TBX is an XML-based family of terminology exchange formats characterised by a hierarchical structure based on the *Terminological Markup Framework* (TMF - ISO 16642:2017).^[3] Specifically, a *terminological data collection* (TDC) is constituted by a set of *concept entries* (CEs), each describing a concept within a specialised domain. Each concept entry includes one or more *language sections* (LSs) that, in turn, include one or more *term sections* (TSs) containing the term designating the concept in that specific language, as well as a set of additional information describing the term itself. In Example 1, we show an example of a TBX file that contains two CE; the first CE identified with the ID "1" has three LSs: English, Spanish, and Italian. The English LS contains two designations for this concept, the same for Spanish, while the Italian LS contains only one designation.

```

<tbx xmlns="urn:iso:std:iso:30042:ed-2" xml:lang="en" style="dct" type="TBX-Min"
  xmlns:min="http://www.tbxinfo.net/ns/min">
  <text>
  <body>
    <conceptEntry id="1">
      <langSec xml:lang="en">
        <termSec>
          <term>mouse pad</term>
        </termSec>
        <termSec>
          <term>mouse mat</term>
        </termSec>
      </langSec>
      <langSec xml:lang="es">
        <termSec>
          <term>soporte para ratón</term>
        </termSec>
        <termSec>
          <term>alfombrilla de ratón</term>
        </termSec>
      </langSec>
      <langSec xml:lang="it">
        <termSec>
          <term>tappetino per mouse</term>
        </termSec>
      </langSec>
    </conceptEntry>
    <conceptEntry id="2">
      ...
    </conceptEntry>
  </body>
</text>
</tbx>

```

Example 1. A TBX file containing a CE with three LSs.

The OntoLex-Lemon model, on the other hand, is characterised by a modular architecture that allows for a detailed and articulated description of the linguistic characteristics of a term. Figure 1 depicts the core elements of the model. Each lexical entry – be it an affix, a word or a multiword expression – is an instance of the class Lexical Entry and is associated to its morphological realisations (class Form) through the relation `ontolex:lexicalForm` as well as to its sense(s) (class Lexical Sense) by means of the `ontolex:sense` property. Lexical sense, conceived as the reification of the `ontolex:denotes` property linking the lexical entry and the concept, can be provided with additional information, such as usage context (`ontolex:usage`), and its semantic relations expressed by means of a particular linguistic vocabulary, e.g., LexInfo.^[4]

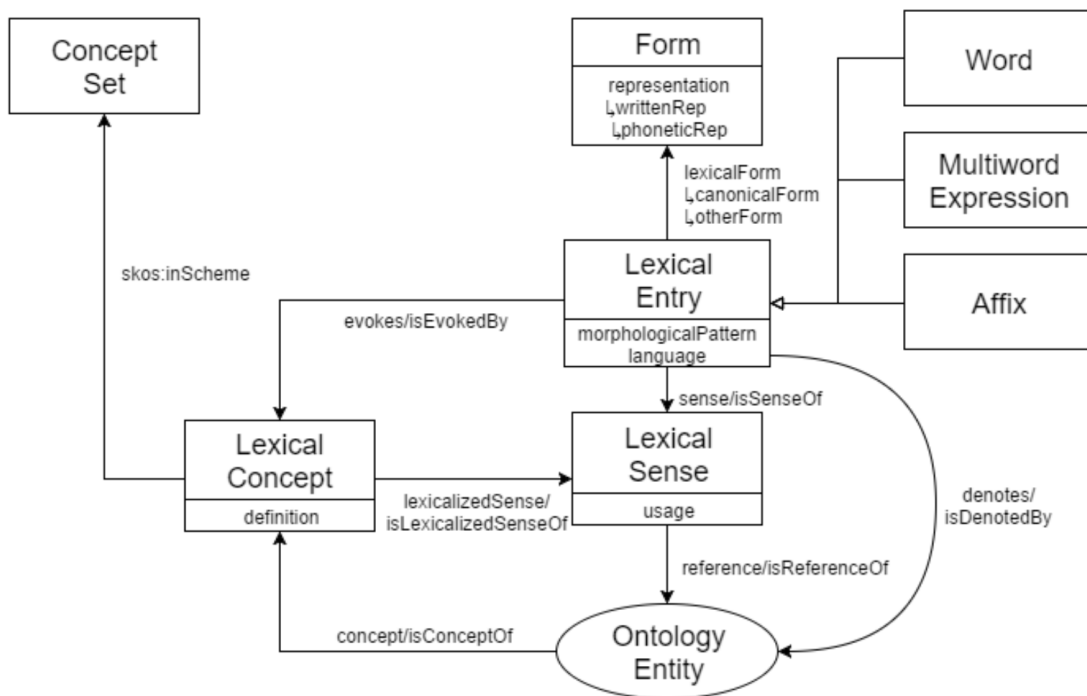


Figure 1. The architecture of the OntoLex-Lemon module.

The concept, viewed as an extra-linguistic entity, can be formally described either in an external domain ontology (already existing or created from scratch) according to the principle known as “semantics by reference” [Buitelaar 2010], or as an instance of the Lexical concept class (see Figure 1), which represents “a mental abstraction, concept or unit of thought that can be lexicalized by a given collection of senses”. In addition to the Ontolex core, other modules have been developed, such as the Decomposition module; the Variation and Translation module; the Syntax and Semantics module; the Linguistic Metadata module.^[5]

To ensure interoperability between these two distinct approaches to data modelling – on one side, the ISO TMF and TBX standards, and on the other side, OntoLex-Lemon and the Semantic Web technologies – we found it necessary to develop a new converter, named LemonizeTBX, which will be the focus of this paper. Although a similar tool was already developed by Cimiano et al (2015), the new converter differs in two fundamental aspects. It is interactive, allowing the terminologist to play an active role in the conversion process. As highlighted by Piccini et al (2023), indeed, moving from a TBX format to an LD structure goes beyond a simple shift from an XML-based to an RDF-based data format. The conversion process involves a shift in perspective: a strictly terminological concept-oriented framework (TBX) is converted into a lexicographical sense-centred standpoint (OntoLex-Lemon). Moreover, as discussed by Bellandi et al (2023b), TBX primarily focuses on facilitating the exchange and administration of terminological resources to maintain coherence and compatibility among terminologists and language experts. Conversely, Ontolex-Lemon is specifically designed to encode lexical data, with the goal of capturing intricate linguistic details and facilitating semantic integration with other RDF datasets.

In addition, the new converter will handle input and output files in the latest versions of TBX and OntoLex-Lemon, as both have undergone revisions over time. In 2019, the TBX version 3.0 was released and published as ISO 30042: 2019, replacing the previous TBX version 2.0 published in 2008 (ISO 30042: 2008). Similarly, the OntoLex-Lemon model, born within the European Monnet project as a result of the integration of the three pre-existing models LingInfo [Buitelaar et al 2006], LexOnto [Cimiano et al 2007], LIR [Montiel-Ponsoda et al 2011], was later extended into the OntoLex-Lemon model by the Ontology-Lexicon Community Group.

This paper is structured as follows. In Section 2, we provide an overview of the main works that address the issue of transforming a terminological resource serialised in a TBX format into a Linguistic Linked Data (LLD) resource serialised

in an RDF OntoLex-Lemon format. In Section 3, we define the key requirements for the development of the new converter LemonizeTBX. Section 4 describes the initial prototype of the converter, and finally, in Section 5, conclusions will be drawn, and future work will be outlined.

2. Transforming TBX: Previous Efforts and On-Going Work

This section provides an overview of previous efforts in transforming a TBX terminological resource into a LLD resource using the OntoLex-Lemon model. We categorised these approaches into two sections: Section 2.1 focuses on the TBX2RDF tool specifically designed for this purpose, and Section 2.2 delves into alternative approaches, discussing the theoretical implications of such transformation.

2.1 An Overview of TBX2RDF Approaches

The application of LD principles to terminological datasets was discussed by Cimiano et al (2015); Siemoneit et al (2015). In their seminal work, the authors present a process involving the mapping of TBX elements and structures into the OntoLex-Lemon ontology, which provides a semantic framework for describing lexical and linguistic information in RDF.^[6] The proposed methodology is implemented as an online service named TBX2RDF that enables the representation of terminological data in a linked and machine-readable format, facilitating their integration with other semantic web resources and applications. In fact, by converting XML data into RDF, it becomes possible to semantically enrich the data, enabling more effective querying, linking, and reasoning across heterogeneous sources on the web. This transformation enhances the accessibility, reusability, and comprehensibility of data, ultimately contributing to the realisation of the Semantic Web's vision of a more interconnected and meaningful web of data. In addition, as the manual creation of links between datasets is costly and therefore not scalable, the authors propose a preliminary attempt to explore approaches for automatic linking between different datasets.

The best practices to model LLD were identified in the masterwork by Cimiano et al (2020b) devoted to the principles behind the idea of semantic web and, in particular, to the issues of reusability of dictionaries and interoperability of resources. However, the treatment of transforming a terminological resource into a LD format often remains confined to a surface-level mapping of elements, lacking deeper consideration for the theoretical implications inherent in these mapping choices.

The design and implementation of TBX2RDF have certainly influenced the research in this area and has given the possibility to create important linguistic linked resources. For example, in Rodriguez-Doncel et al (2018), the authors describe a terminology created in a half-automated process, where terms and natural language definitions have been extracted and integrated from different lexical sources and mapped in a supervised process. In particular, the paper analyses the process and the methodologies adopted in the automatic conversion into TBX of linguistic resources together with their semantic enrichment.

A similar, but somewhat inverse process, was described by Speranza et al (2020) who propose a proof-of-concept for an automated methodology to convert linguistic resources into TBX resources using the Italian Linguistic Resource for the Archaeological domain. A conversion tool is designed and implemented to support the creation of ontology-aware terminologies, enhancing interoperability and facilitating the sharing of language technologies and datasets. By ontology-aware terminologies, the authors intend the approach that relies on a precise conceptual mapping of linguistic and semantic information to TBX elements and the enrichment of these elements by means of semantic information, which may be useful in several applications to create a cloud of interoperable and interconnected terminologies, directly linked to both already existing ontologies and new developed ones.

In Declerck et al (2020); di Buono et al (2020), the authors discuss the Prêt-à-LLOD project^[7] which aims at increasing the uptake of language technologies by exploiting the combination of LD and language technologies to create ready-to-use multilingual data. In particular, they present a new approach called Terme-à-LLOD devoted to simplifying the task of converting a TBX terminological resource into a LLD one.

Another example of the organisation of terminological data concerning the domain of raw material and mining was

presented by Kitanovic et al (2021). The authors introduce the development of a termbase for the field of mining engineering, the transformation from a custom scheme into the TBX, and then a further analysis aiming at compatibility with the LD approach.

In Chozas and Declerck (2022), the authors explore the use of the OntoLex-Lemon model and suggest some extensions, for achieving a declarative encoding of relations between multilingual expressions contained in terminologies. In particular, the authors link their proposal to the previous LIDER project^[8] that was already concerned with mapping TBX to RDF, with the goal of transforming and publishing terminologies as LD.

All the previous works have set an important base for the development of future tools. Nevertheless, there are at least two issues that were overlooked: first, the lack of a theoretical analysis about the implications of such a transformation – can we really use a lexicographic model for representing a terminological resource and vice versa? – second, the fact that none of the most recent works have implemented the tools on the latest TBX standard (ISO 30042: 2019).

In the following Section, we present the papers that have raised these questions together with alternative approaches to the mapping of TBX files into other formats.

2.2 Open Issues and Other Approaches

Reineke and Romary (2019) were among the pioneers in questioning the theoretical soundness of converting a terminological resource from TBX to RDF. In their article, the authors discuss the limitations of the outcomes generated by TBX2RDF due essentially to the fact that a semasiological (word-to-sense) model such as OntoLex cannot be naturally mapped onto the concept-to-term model of TBX. Consequently, they present a comparison between SKOS and TBX, which, conversely, in their view, prove to be more interoperable given their shared foundation in a concept-oriented perspective. As discussed in detail by Bellandi et al (2023b), the choice of whether and how to represent senses from a data model that does not inherently incorporate senses is indeed one of the most important issues in these data serialisation procedures. We will return to this point in Section 3.1 to illustrate different approaches to addressing this challenge.

A multilevel contrastive analysis between TBX and OntoLex-Lemon was presented by Bellandi et al (2023a); Piccini et al (2023), taking into account the technological, theoretical, methodological dimensions, as well as the user perspective and application of the two data models. The authors provide a detailed discussion about the issues of the already available tools for this type of conversion, and the need for a deeper understanding of both models and the ability to reconcile the differences in their structures and semantics during the conversion process.

Some other researchers provide a different perspective on the possibility of employing alternative serialisations and representations. For example, Suchowolec et al (2019) discuss the limitations of TBX in terms of the three levels (concept, language, and term) and the use of SKOS as a potential format to describe the amount of available relation types. A different point of view is given by Di Nunzio and Vezzani (2021) who put forward the hypothesis that, if terminological data are appropriately modelled at an abstract level, they can be exported or exposed in any format, while preserving both the linguistic dimension and its relationship to the conceptual dimension.

3. Analysis of Requirements

Before delving into the core content of the paper, we define the main requirements underpinning the design and the implementation of the converter.

3.1 Theoretical Flexibility

The terminology community encompasses a range of heterogeneous theoretical and methodological perspectives.^[9] In a nutshell, three main theoretical approaches can be distinguished:

1. The concept has priority over the term that is a linguistic designation with the function of naming the concepts. This would be the view of classical terminology that would relegate the contextual interpretation

of the sense of the term to a secondary role.

2. The concept designated by the term is flattened on the sense of the term itself, and no difference is made between the linguistic and conceptual dimensions of the terminological analysis.^[10]
3. The concept and the sense are two distinct entities that both need to be taken into consideration, in order not to turn terminology into specialised lexicography. Terminology, indeed, is considered as a “twofold science” characterised by two complementary and necessary dimensions of analysis, namely the conceptual and the linguistic [Costa 2013]; [Santos and Costa 2015].

The flexibility of the OntoLex-Lemon model makes it possible to satisfy all these three main theoretical approaches. Consequently, as we shall see in more detail in Section 4, the converter will make it possible to:

22

- bypass the Lexical Sense class and directly link the lexical entry to the concept it designates;
- omit the conceptual dimension and limit itself to the description of the sense via the Lexical Sense class;
- describe both dimensions. As far as the conceptual dimension is concerned, the converter is agnostic with respect to the specific ontology language used for representing concepts. It is up to the terminologist to decide if using SKOS, OWL, or mapping the concepts to a given ontology.

3.2 Reusability

The LD paradigm strongly advocates for the reuse of existing vocabularies. In line with this principle, the new converter will provide terminologists with the flexibility to choose the data categories they deem more suitable, thus facilitating:

23

- Interoperability: Terminologists might feel the need to align their resources with widely accepted standards and ontologies.
- Flexibility for domain-specific needs: Terminologists may find it necessary to effectively address the specific requirements and nuances for data categorization set by their specialised domain.
- Adaptability for evolving standards: Terminological needs and standards can evolve over time. Adaptability in this sense is ensured by the active involvement of the users in the conversion process. They can decide the latest standards in terminology to conform with or the new best practices emerged within the LD community.

3.3 Knowledge Extraction

When employing TBX to describe terminological data, researchers may encounter constraints dictated by their chosen TBX dialect.^[11] Each dialect has its own unique set of data categories designed to describe the entries, and the richness of this set directly influences the granularity of the description of each entry. Despite the availability of a large shared set of categories, challenges arise when terminographers lack a specific data category to describe a specific information. Let us suppose that terminographers need to add information about the etymology of a term. In such cases, they have no choice but to use the `<note>` field to store this information. Notably, the strategy adopted by TBX2RDF is to use annotation properties, such as `rdfs:comment`, `rdfs:label`, and so on.^[12] Nevertheless, our perspective underscores the importance of employing automatic methods to first analyse these unstructured text sources, then identify structured semantic information embedded within the text, and subsequently organise, and store it by means of the appropriate relations within OntoLex-Lemon. Returning to our example regarding etymology, OntoLex-Lemon provides the `LemonEty`^[13] module specifically designed to accurately represent such information. This proactive strategy ensures that valuable semantic information, initially confined to notes, is systematically integrated into the structured ontology, thus fostering a more comprehensive and enriched representation of terminological knowledge.

24

3.4 Data Enrichment

This requirement involves the explication of information that is implicitly conveyed by the hierarchical structure of TBX itself, and explicitly written in unstructured content such as note fields of a concept entry in TBX.

25

3.4.1 Deductive Rules Exploitation

The hierarchical structure of TBX implies relationships that can be lost in the conversion from TBX to OntoLex-Lemon, if not properly managed. These relationships mainly include synonymy and equivalence. In the first case, terms described within the same LS belonging to the same concept entry are assumed to be synonymous without the necessity of adding an explicit synonymy relationship among them. Likely, in TBX, terms in different LSs, grouped within the same concept entry, are equivalents. By employing deductive rules, our converter LemonizeTBX enables the terminographer to explicitly define these relations in the new resource, suggesting the appropriate links (e.g. `lexinfo:synonym` for synonymy and `vartrans:translatableAs` or `lexinfo:translation` for linguistic equivalence among terms).

26

3.4.2 TBX Data Enrichment

If the extraction of knowledge from unstructured notes proves successful (see Section 3.3), an additional step may involve enhancing the original TBX by incorporating the newfound information, in a sort of hermeneutic circle. This process goes beyond a naive addition of data, as it aims to complement the relationships stored within the TBX file, whether implicit or explicit.

27

While this proposition stands as a theoretically sound alternative, practical implementation considerations arise when modifying the TBX structure within the framework of the latest ISO standard. The pipeline of this process becomes easier by leveraging pre-established data categories available in DatCatInfo.^[14] This resource is a data category repository collecting data categories specifically developed for terminology work. The strategic use of these existing categories facilitates seamless integration and compliance to standard specifications. Conversely, should there be a need to introduce a novel data category, a more elaborate solution is required to document and comply with the requirements of the ISO standards. This expanded documentation process ensures that any additions align seamlessly with the established standards, maintaining the integrity and compatibility of the enriched TBX structure.

28

4. LemonizeTBX: Towards a TBX to OntoLex-Lemon Converter Requirements

In this Section, we present the design and the current development of LemonizeTBX, the prototype for converting resources from TBX into OntoLex-Lemon. Following the considerations mentioned above, we started to develop an interactive and highly customizable converter designed to accommodate the theoretical framework of the user conducting the conversion, be they terminologists, translators, or lexicographers. Given the dynamic nature of the converter, it also became crucial to develop a user interface that enables the coordination and monitoring of the conversion process. From a technological standpoint, we designed a software architecture based on the back-ends-for-front-ends pattern: a server-side component implementing the logic of conversion processes, and a client-side application representing the interface through which a user interacts and manages these processes.

29

In the following, Section 4.1 is devoted to providing the implementation details of the server component (referred to as back-end from here on), while Section 4.2 describes the user interface through an example of conversion. Finally, Section 4.3 illustrates the current limitations of our prototype, and provides some general considerations.

30

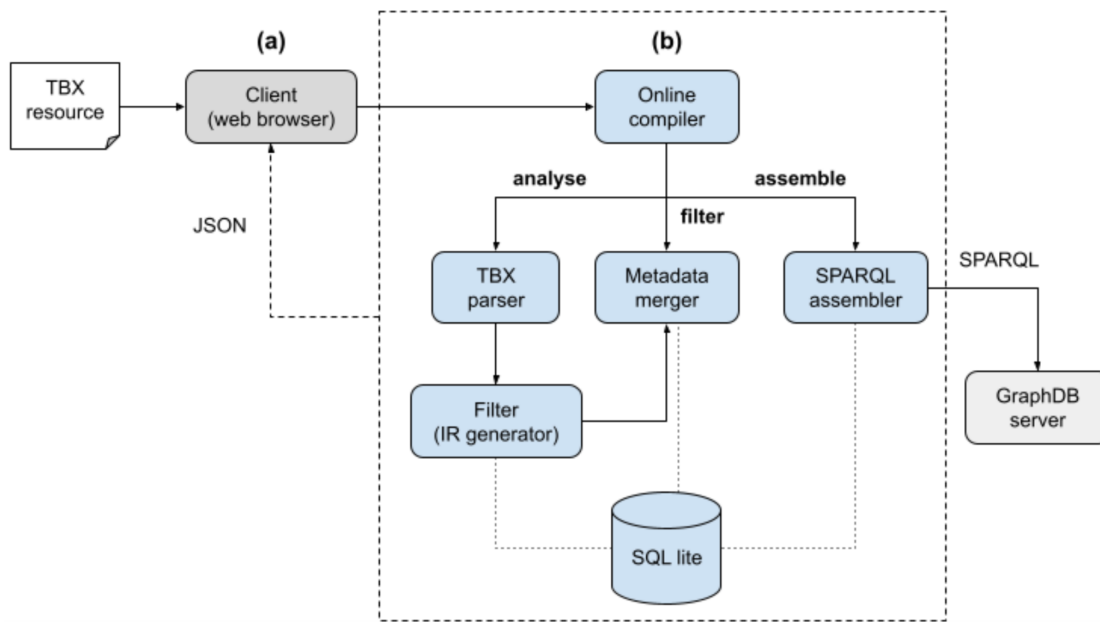


Figure 2. The architecture of LemonizeTBX converter: the front-end discussed in Section 4.2 and the back-end discussed in Section 4.1.

4.1 The Back-end Architecture

The architecture, namely an Online Compiler, translates a TBX source into RDF triples according to the OntoLex-Lemon model, going interactively through three main phases: analysis, filtering, and assemblage, as depicted in Figure 2. Through a Web client (described in the next section), a user can upload a TBX resource that is ingested by the Online Compiler that starts the process:

31

- In the first phase, analysis, the Online Compiler routes the TBX file to the TBX parser component, which is in charge of analysing the XML input – potentially written in different TBX dialects (min, basic, core) – and processing the resource in order to get an *intermediate representation* (IR) of the information contained. IR is designed to fulfil two objectives: perform a partial conversion of entities (lexicons, concepts, terms, and senses) into their RDF-equivalent series of triples, and convey the resource's metadata information (such as the number of languages and terms) in JSON format. This format facilitates user interaction during the filtering and assemblage phases. Please note that the conversion at this stage is not considered final, as the ultimate coalescing phase will necessitate additional input from the user. During this phase, the resource is converted without making any assumptions about the final output.
- The second phase, filtering, enables the user to query and inspect the resource, allowing them to define a personalised filter for selecting and potentially enriching the data. All interactions with the client are conducted by the Metadata Merger, which queries the database and returns a JSON response to the web client.
- Subsequently, the assemblage phase, starting from the filtered data, constructs the OntoLex-Lemon lexica. This process involves processing languages, concepts, and terms, and serialising them as RDF triples in accordance with the OntoLex-Lemon data model.

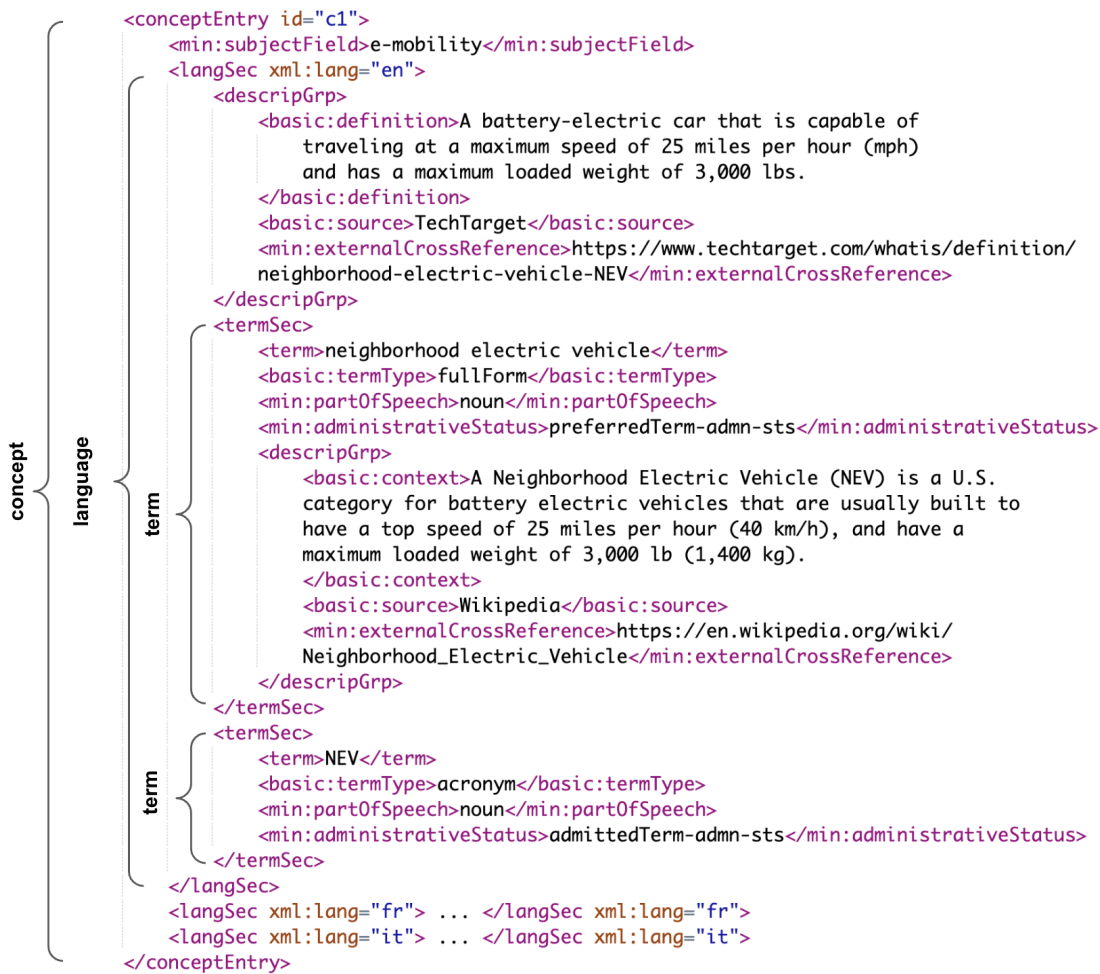


Figure 3. The hierarchical structure of a TBX entry (<conceptEntry>). It has a set of language sections (<langSec>), and for each of them, a set of terms that designate the concept for that language (<termSec>). The example is about the concept of a *neighborhood electric vehicle* (c1), that is designed by the English terms *neighborhood electric vehicle*, and *NEV*. Please find an accessibly formatted version of this annotated code sample in the Appendix.

4.2 The Front-end Prototype

Throughout the design and development of the converter, we endeavoured to envision the interaction between the user and the converter, guided by the requirements and considerations outlined in Section 3. In this section, we will illustrate the user interface using a brief example of TBX (basic). For simplicity, our focus will be on the concept of *neighborhood electric vehicle*. Figure 3 displays the XML code related to the English language section of this concept. Specifically, the fragment presents a concept identified as “c1” within the e-mobility field, along with an English language section containing the definition for that concept. Within the English language section, there are two terms associated with “c1”: the preferred full form *neighborhood car vehicle* and the accepted acronym *NEV*. Each term includes specific information, such as morphology, usage contexts, and external references. Subsequently, we will explore incrementally how each part of the TBX input is converted into RDF triples according to the OntoLex-Lemon model.

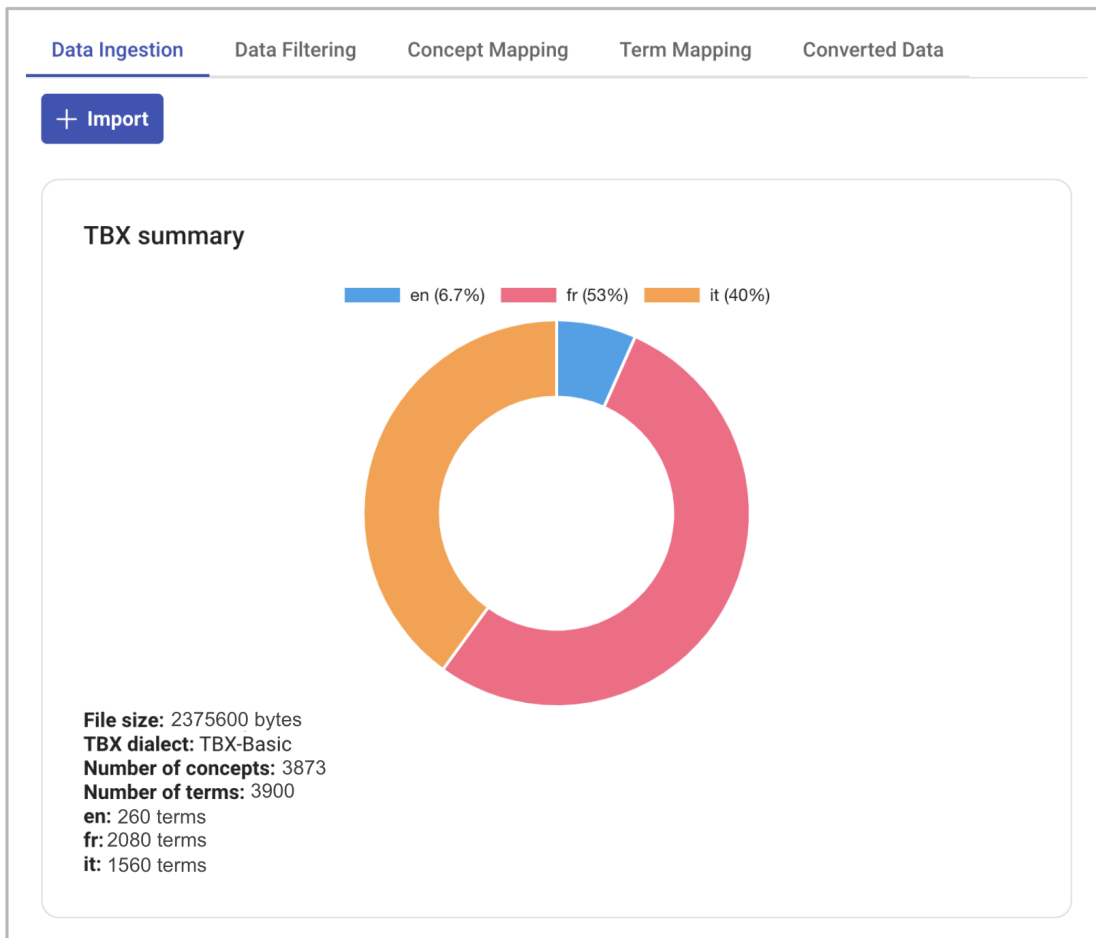


Figure 4. Representation of a quantitative summary of the imported TBX resource.

The interface divides the conversion process into five key phases:

33

1. Data ingestion: makes it possible to upload the TBX file to be converted, and verifies that it is written in one of the three TBX dialects (basic, min, core); validate it according to the three dialects (basic, min, core);
2. Data filtering: allows users to specify the data they wish to convert;
3. Concept mapping: permits users to define how filtered concepts should be converted;
4. Term mapping: lets users specify the conversion approach for terms associated with filtered concepts;
5. Data conversion: displays the converted data by querying the triple store containing the conversion.

Figure 4 illustrates the initial phase of the conversion process. Users can upload a TBX file, and the system parses the data, constructing an internal intermediate representation (IR) as outlined in Section 4.1. It is important to note that the input TBX file is assumed to be valid and well-formed; otherwise, the system will generate an error, a topic we will delve into further in Section 4.3. The system provides the user with a summary of the ingested data, presenting information such as file size, the public TBX dialect used, the count of concepts, and the number of terms per concept and language. This summary proves valuable for obtaining a quantitative overview of the file slated for conversion.

34

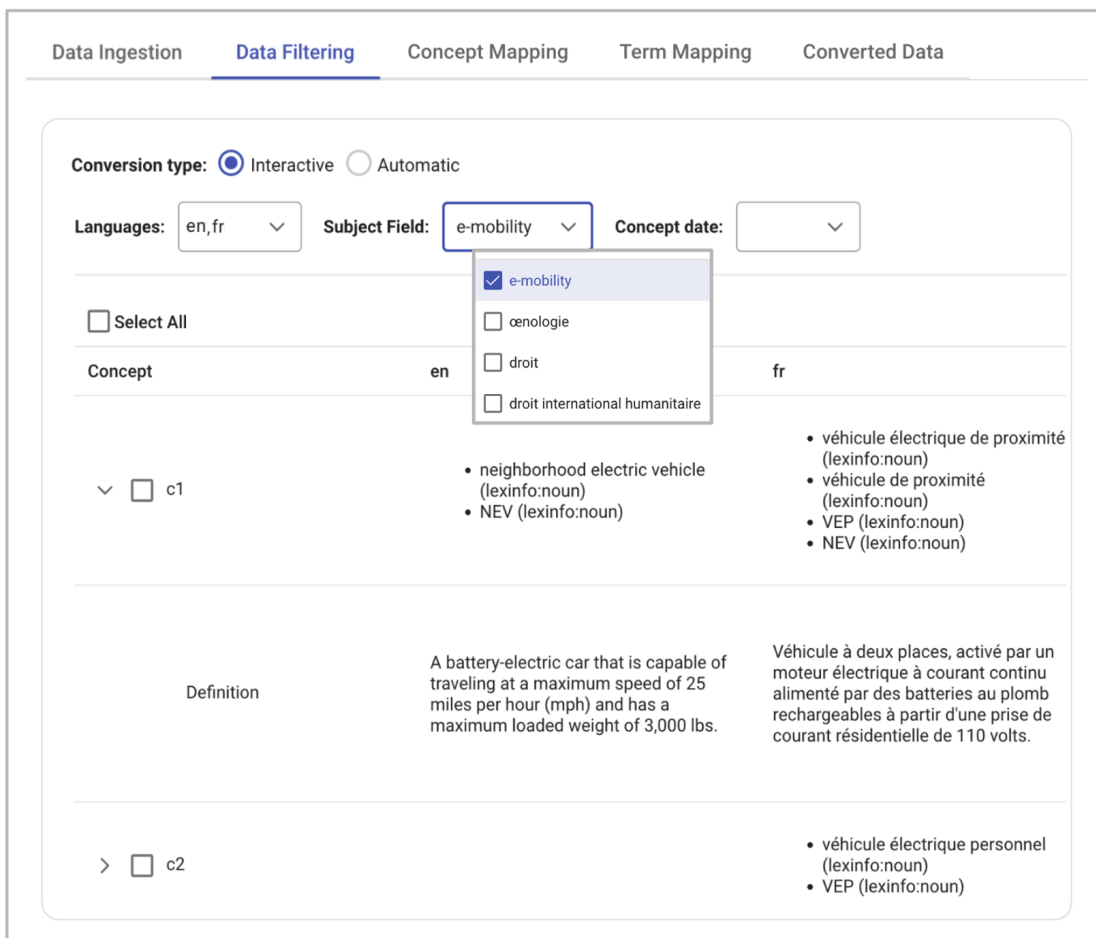


Figure 5. TBX data filtering interface.

At this point, the user has the option to either proceed with an automatic file conversion using a default mapping or interactively determine the conversion of terminological entities on a case-by-case basis, as illustrated in Figure 5. The system presents TBX data in a table featuring a column for concepts and additional columns for each language in the TBX file. Each column includes all the terms designating the respective concepts, along with their parts of speech. Rows in the table, such as the one for concept c1 in Figure 5, can be expanded to reveal the definition of the concept in natural language across different languages. This capability stems from the TBX example used, where concept definitions are provided at the language level (see the <descripGrp> end-tag in Figure 3). In instances where the definition is specified at the concept level, the system directly presents it in the concept column. The table results can be filtered based on two parameters, which can also be used in combination: concepts created on specific dates (concept date filter) and concepts associated with particular topics (subject field filter). Additionally, users can opt to convert terms only in certain languages using the language filter. Subsequently, users can select the concepts they wish to convert by checking the checkbox in the corresponding column.

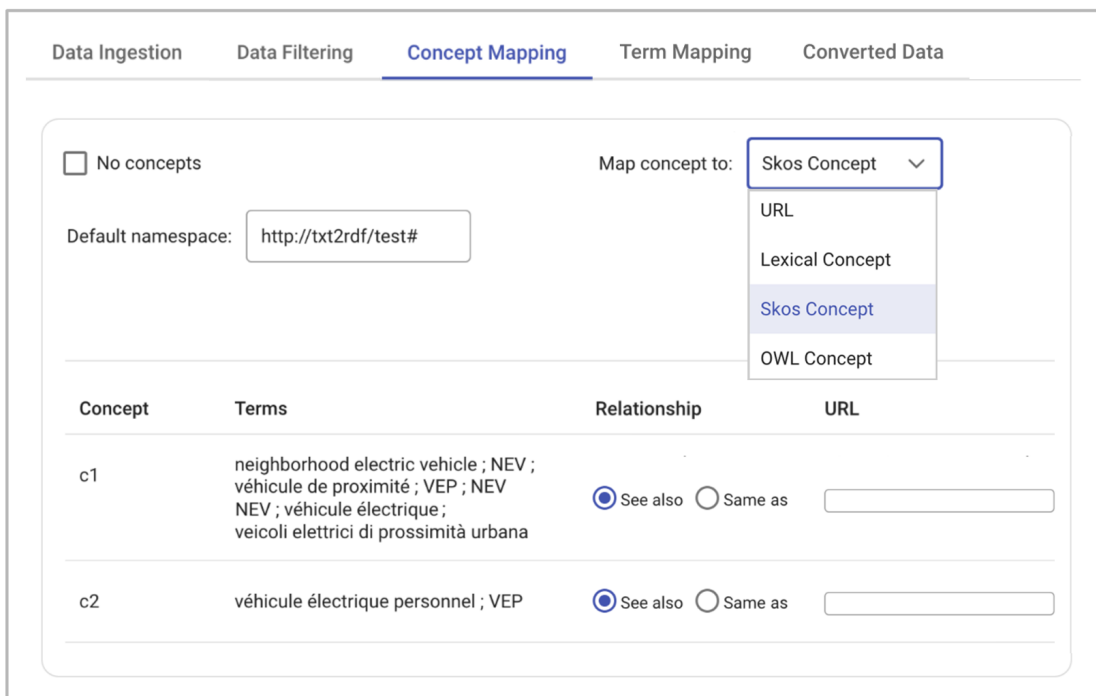


Figure 6. TBX concepts mapping interface.

After selecting the concepts, users can determine how to convert them, as depicted in Figure 6. It is also an option to convert only the language part, i.e., the terms designating the selected concepts. There are four conversion possibilities:

36

1. *URL*: the concepts are not created as ontological entities, but they are defined by URLs specified by the user;
2. *Lexical Concept*: the system maps concepts to the class *Lexical Concept*, representing a mental abstraction, concept, or unit of thought lexicalized by a collection of terms;
3. *SKOS Concept* (default choice for automatic conversion): concepts are converted into SKOS classes. A detailed example will be presented below;
4. *OWL Concept*: the system generates as many OWL classes as selected concepts.

Additionally, users can specify whether, in relation to a concept “C” already defined in another ontology, that class represents it exactly (`owl:sameAs`) or is related in some way (`rdfs:seeAlso`). Concerning the SKOS choice, the concepts (`<conceptEntry>`), are converted by means of the SKOS ontology, according to Reineke and Romary (2019). Subject fields correspond to SKOS concept schemes, while concepts are mapped to SKOS concepts. The membership of concepts to their subject fields is formalised through the SKOS `textit{inScheme}` relationship. The SKOS `definition` property of a concept represents the definition of that concept provided by the TBX resource, whether the definition is given at the concept level or at the language level.

37

In Example 2, we provide a sample of the RDF Turtle serialisation produced by the tool:

38

```

:cl a skos:Concept ;
  skos:prefLabel "c1"@en ;
  skos:inScheme :sbjf_1 ;
  skos:definition "A battery-electric car that is capable of traveling at a maximum
    speed of 25 miles per hour (mph) and has a maximum loaded weight of 3,000
    lbs."@en ;
  skos:definition "véhicule à deux places, activé par un moteur électrique à courant
    continue alimenté par des batteries au plomb rechargeables à partir d'une prise
    de
    courant résidentielle de 110 volts."@fr .
:sbjf_1 a skos:ConceptScheme ;
  skos:prefLabel "e-mobility"@en .

```

Example 2. Sample Turtle serialisation produced by LemonizeTBX for definitions.

For each <langSec> element in original TBX resource, related OntoLex-Lemon lexica are created. Taking as an example the case in Figure 3, English, French, and Italian lexica are defined, and the terms are created as entries of the suitable lexicon. The Turtle serialisation produced by the tool is provided in Example 3:

39

```

:enLex a lime:Lexicon ;
  lime:language "en" ;
  lime:entry :t1, :t2 .
:frLex a lime:Lexicon ;
  lime:language "fr" ;
  lime:entry :t3 .
:itLex a lime:Lexicon ;
  lime:language "it" ;
  lime:entry :t4, :t5 .

```

Example 3. Sample Turtle serialisation produced by LemonizeTBX for three lexica.

Finally, users have to determine how to convert the terms. The interface depicted in Figure 7 makes it possible to establish:

40

1. whether lexical senses of the terms should be created;
2. what semantic relations among terms need to be represented;
3. how to handle terms with the same character sequence but different meanings.

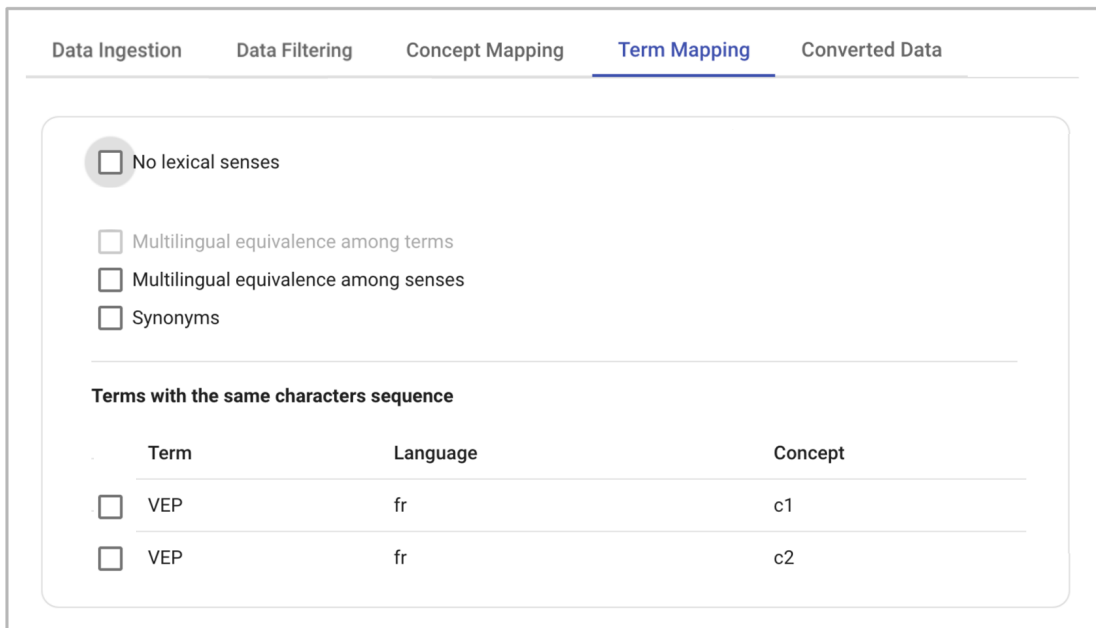


Figure 7. TBX terms mapping interface.

For point one, the terms within the <termSec> entity are represented as lexical entries in the OntoLex-Lemon model. Each term is mapped to the Lexical Entry class, without specifying its particular type (word or multi-word), and it is represented as a canonical form, typically a lemma, of that lexical entry. Following the *semantics by reference* paradigm of OntoLex-Lemon, the meaning of a lexical entry is specified by referring to the created SKOS concept that represents its meaning.

41

The default conversion process generates a lexical sense for each lexical entry and links it to the suitable concept by means of the `ontolex:reference` property. In the event the user opts not to represent the lexical sense of lexical entries (option shown in Figure 7), the converter links the term directly to the concept by means of the `denote` property. Given that the model lacks a comprehensive set of linguistic categories, it relies on Lexinfo vocabulary.^[15] Consequently, morphological information, such as part of speech, and other morphological traits is associated with the forms, while usage context, term type, and administrative status are associated with the senses, according to the Lexinfo schema. If the definition in the <langSec> has a source or/and an external reference, the reification mechanism^[16] would be employed in order to represent the source and the reference of the concept definition, using Dublin core source, and RDF `seeAlso` properties, respectively. Similarly, this mechanism can be applied to represent examples of term usage with the `usage` property. Example 4 demonstrates the conversion related to the term *neighborhood electric vehicle*:

42

```

:t1 rdf:type ontolex:LexicalEntry ;
  lexinfo:partOfSpeech lexinfo:noun ;
  lexinfo:normativeAuthorization lexinfo:preferredTerm ;
  lexinfo:termType lexinfo:fullForm ;
  ontolex:canonicalForm :t1_cf ;
  ontolex:sense :t1_sense .
:t1_cf rdf:type ontolex:Form ;
  ontolex:writtenRep "neighborhood electric vehicle"@en .
:t1_sense rdf:type ontolex:LexicalSense ;
  skos:definition [
    rdf:value "A battery-electric ..." ;
    dct:source "TechTarget" ;
    rdfs:seeAlso <https://www.techtarget.com/whatis/definition/
      neighborhood-electric-vehicle-NEV> ] ;
ontolex:usage [
  rdf:value "A Neighborhood Electric Vehicle is a U.S. category for ..." ;
  dct:source "Wikipedia" ;
  rdfs:seeAlso <https://en.wikipedia.org/wiki/Neighborhood_Electric_Vehicle> ] .

```

Example 4. Converting *neighborhood electric vehicle* with LemonizeTBX.

Concerning point two, the semantic relations among terms, synonymy and translation can be identified and represented in OntoLex-Lemon. In particular, if the appropriate checkboxes in Figure 7 are selected, the system retrieves all the terms in different languages belonging to the same `<conceptEntry>` element and creates suitable translation relations among them^[17]. Similarly, the system retrieves the terms of each language that share the same meaning (i.e., design the same concept) and converts them as synonyms. It is important to note that this is only feasible if the user opts to create lexical senses, as the Lexinfo synonymy relationship links senses together. The default behaviour in the case of automatic conversion is to refrain from representing any semantic relation.

43

Finally, as for terms with the same character sequence but different meanings (see point three), all terms designating more than one concept are listed in a table. The user can choose which terms to represent as polysemic entries (multiple lexical senses are created for each term), and which terms to represent as homonyms (multiple distinct lexical entries are created). The default behaviour in the case of automatic conversion is the latter.

44

4.3 Considerations

The development of the converter is still in its early stages, and currently, only the default behaviour, outlined in the preceding sections, has been implemented. In this Section, we briefly address the limitations of the current implementation and how they manifest in the converted data.

45

The converter presupposes that the TBX file slated for conversion is both syntactically and semantically valid. This implies that the file must be well-formed and the textual values within the XML tags should align semantically with the intended representation of the tags. Specifically, it assumes accurate reporting of data category values to ensure proper mapping with the Lexinfo vocabulary. Currently, only public TBX dialects are handled^[18], and incorporating a private dialect would necessitate adjustments to the converter code, particularly in updating the parser and IR generator depicted in Figure 2. Introducing specific data categories might also require modifications to the user interface code. A pivotal aspect of the conversion process is the decision on whether or not to represent lexical sense. As emphasised in Section 3, the importance of sense may vary, and it could either not play a central role or be flattened onto the concept. However, in the OntoLex-Lemon model, the sense is deemed central, with semantic relations formally defined as links

46

between lexical senses.^[19] Choosing not to represent lexical sense limits the converter's ability to explicitly define semantic relations, as illustrated by the interface in Figure 7.

An alternative approach could involve mapping semantic relations (such as, for example, hypernymy, meronymy, etc.) with relations between the concepts (such as isA, partOf, etc.) designated by the terms to which the appropriate senses would correspond. This aligns with perspectives that do not distinguish between sense and concept, but deviates from the theoretical viewpoint where senses and concepts are considered distinct entities, albeit intimately related.

47

5. Conclusions

The paper introduces LemonizeTBX, a prototype software to enhance interoperability between the ISO TBX standard, used for representing terminological resources and the OntoLex-Lemon model, adopted to represent lexicographic resources. The converter aims to bridge the gap between terminological and lexicographical frameworks, emphasising a shift from a concept-oriented to a sense-centred standpoint. This change in perspective underscores the need for an interactive tool to engage terminologists in the conversion process. The interactivity concerns choices that users need to make when there is no direct mapping from TBX to OntoLex-Lemon. For instance, deciding whether and how to link lexical entries directly to concepts or how to add missing data categories. The converter is designed to support an advanced process to enrich the data by extracting implicit and explicit information from unstructured content in TBX entries.

48

The implementation of LemonizeTBX is currently in its early stages, featuring some default behaviour. For instance, it assumes syntactic and semantic validity in the input TBX files, it primarily focuses on representing lexical senses, leaving room for potential alternative approaches, such as mapping semantic relations to concept relations.

49

However, as discussed in the preceding Section 4.3, the work conducted thus far has revealed that the challenges posed by the conversion process are fundamentally theoretical, raising questions that deserve discussion within the community. This discussion is crucial to provide a solid foundation for the development of a customised OntoLex-Lemon module for terminology science tailored to the specificities of this discipline.^[20]

50

Appendix


```

<conceptEntry id="cl">
  <min:subjectField>e-mobility</min:subjectField>
  <langSec xml:lang="en">
    <descripGrp>
      <basic:definition>A battery-electric car that is capable of
        traveling at a maximum speed of 25 miles per hour (mph)
        and has a maximum loaded weight of 3,000 lbs.
      </basic:definition>
      <basic:source>TechTarget</basic:source>
      <min:externalCrossReference>https://www.techtarget.com/whatis/definition/
        neighborhood-electric-vehicle-NEV</min:externalCrossReference>
    </descripGrp>
    <termSec>
      <term>neighborhood electric vehicle</term>
      <basic:termType>fullForm</basic:termType>
      <min:partOfSpeech>noun</min:partOfSpeech>
      <min:administrativeStatus>preferredTerm-admn-sts</min:administrativeStatus>
      <descripGrp>
        <basic:context>A Neighborhood Electric Vehicle (NEV) is a U.S.
          category for battery electric vehicles that are usually built to
          have a top speed of 25 miles per hour (40 km/h), and have a
          maximum loaded weight of 3,000 lb (1,400 kg).
        </basic:context>
        <basic:source>Wikipedia</basic:source>
        <min:externalCrossReference>https://en.wikipedia.org/wiki/
          Neighborhood_Electric_Vehicle</min:externalCrossReference>
      </descripGrp>
    </termSec>
    <termSec>
      <term>NEV</term>
      <basic:termType>acronym</basic:termType>
      <min:partOfSpeech>noun</min:partOfSpeech>
      <min:administrativeStatus>admittedTerm-admn-sts</min:administrativeStatus>
    </termSec>
  </langSec>
  <langSec xml:lang="fr"> ... </langSec xml:lang="fr">
  <langSec xml:lang="it"> ... </langSec xml:lang="it">
</conceptEntry>

```

Example 5. Editor's addition of Figure 3 in plain text. This version lacks the brackets which vertically indicate the extent of each XML element in the code sample.

Notes

[1] See <https://www.iso.org/standard/62510.html>

[2] Another data model widely used in the world of semantic web for sharing and publishing lexical information as LD is the *Simple Knowledge Organization System* (SKOS). The latter features the same concept-centric structure as TBX (for a comparison between TBX and SKOS see [Reineke and Romary 2019]).

[3] See <https://www.iso.org/standard/56063.html>

[4] LexInfo is an ontology that provides data categories for the OntoLex-Lemon model.

[5] For a description of the above-mentioned modules see: <https://www.w3.org/2016/05/ontolex/>

[6] See <https://www.w3.org/2015/09/bpmlod-reports/multilingual-terminologies/> for details.

[7] See <https://cordis.europa.eu/project/id/825182>

[8] See <https://www.multilingualweb.eu/projects/lider>

[9] For a complete and recent overview of the different theoretical perspectives on terminology science, see Faber and L'Homme (2022).

[10] As underlined by Rastier: “le concept est le signifié d'un mot dont on décide de négliger la dimension linguistique” [Rastier 1995, 55].

[11] The TBX format can be implemented via both public and private dialects. The public dialects are: TBX-Core, TBX-Min, and TBX-Basic. For a detailed description of these dialects, see: <https://www.tbxinfo.net/tbx-dialects/>

[12] Please see https://www.w3.org/TR/rdf-schema/#ch_properties for a complete list.

[13] Please, see Khan (2018).

[14] See <https://datcatinfo.termweb.eu/>

[15] LexInfo is an ontology that provides data categories for the `\textit{lemon}` model. See <https://lexinfo.net/>

[16] The reification is a mechanism allowing us to write RDF triples about RDF triples. In our case, we could specify both the source and the link of concept definitions.

[17] For example, referring to Figure 6, all the terms referring to concept `c1` are in a translation relationship with each other.

[18] For efficiency reasons, the converter still does not handle TBX files that are too large, such as IATE exports.

[19] In the terminology model described in ISO 30042:2019, the entity `lexical_sense` is not defined.

[20] At the time of writing this article, the W3C OntoLex group is discussing the possibility of introducing such a module https://www.w3.org/community/ontolex/wiki/Terminology\#Use_Cases

Works Cited

Bellandi et al 2023a Bellandi, Andrea, Di Nunzio, Giorgio Maria, Piccini, Silvia, and Vezzani, Federica. (2023a) “From TBX to Ontolex Lemon: Issues and Desiderata” in *Proceedings of the 2nd International Conference on Multilingual Digital Terminology Today* (MDTT 2023), Volume 3427 of *CEUR Workshop Proceedings*, Lisbon, Portugal. CEUR. ISSN: 1613-0073. Available at: <https://ceur-ws.org/Vol-3427/paper4.pdf> (Accessed: 09 April 2024).

Bellandi et al 2023b Bellandi, Andrea, Di Nunzio, Giorgio Maria, Piccini, Silvia, and Vezzani, Federica. (2023b) “The Importance of Being Interoperable: Theoretical and Practical Implications in Converting TBX to OntoLex-Lemon” in *Proceedings of the 4th Conference on Language, Data and Knowledge*, pages 646–651, Vienna, Austria. NOVA CLUNL, Portugal. Available at: <https://aclanthology.org/2023.ldk-1.70> (Accessed: 09 April 2024).

Bosque-Gil et al 2015 Bosque-Gil, Julia, Gracia, Jorge, Aguado-de Cea, Guadalupe, and Montiel-Ponsoda, Elena. (2015) “Applying the OntoLex Model to a Multilingual Terminological Resource” in *The Semantic Web: ESWC 2015 Satellite Events, Lecture Notes in Computer Science*, Cham, pp. 283–294. Springer International Publishing. Available at: https://doi.org/10.1007/978-3-319-25639-9_43 (Accessed: 09 April 2024).

Buitelaar 2010 Buitelaar, Paul. (2010) “Ontology-based semantic lexicons: mapping between terms and object descriptions” In A. Gangemi, A. Lenci, A. Oltramari, C.-r. Huang, L. Prevot, and N. Calzolari (Eds.) *Ontology and the Lexicon: A Natural Language Processing Perspective, Studies in Natural Language Processing*, pp. 212–223. Cambridge: Cambridge University Press. Available at: <https://doi.org/10.1017/CBO9780511676536.013> (Accessed: 09 April 2024).

Buitelaar et al 2006 Buitelaar, Paul, Declerck, Thierry, Anette Frank, Racioppa, Stefania, Kiesel, Malte, Sintek, Michael, Engel, Ralf, Romanelli, Massimo, Sonntag, Daniel, Loos, Berenike, Micelli, Vanessa, Porzel, Robert, and Cimiano, Philipp. (2006) “LingInfo: Design and Applications of a Model for the Integration of Linguistic Information in Ontologies” in *Proceedings of the OntoLex Workshop at LREC 2006*. Available at: https://www.aifb.kit.edu/images/9/9f/2006_1233_Buitelaar_LingInfo_Desig_1.pdf (Accessed: 09 April 2024).

Chozas and Declerck 2022 Chozas, Patricia Martin and Declerck, Thierry. (2022) “Representing Multilingual Terminologies with OntoLex-Lemon (short paper)” in *Proceedings of the 1st International Conference on Multilingual Digital Terminology Today* (MDTT 2022), Volume 3161 of *CEUR Workshop Proceedings*, Padua, Italy. CEUR. ISSN: 1613-0073. Available at: <https://ceur-ws.org/Vol-3161/short1.pdf> (Accessed: 09 April 2024).

Cimiano et al 2007 Cimiano, Philipp, Haase, Peter, Herold, Matthias, Mantel, Matthias, and Buitelaar, Paul. (2007) “Lexonto: A model for ontology lexicons for ontology-based nlp” in *Proceedings of the OntoLex07 Workshop held in conjunction with ISWC'07*. Available at: <https://www.aifb.kit.edu/web/Inproceedings1584> (Accessed: 09 April 2024).

- Cimiano et al 2015** Cimiano, Philipp, McCrae, John P., Rodríguez-Doncel, Víctor, Gornostay, Tatiana, Gómez-Pérez, Asunción, Siemoneit, Benjamin and Lagzdins, Andis. (2015) "Linked terminologies: applying linked data principles to terminological resources" in *Proceedings of the eLex 2015 Conference*, pp. 504–517. Available at: https://elex.link/elex2015/proceedings/eLex_2015_34_Cimiano+etal.pdf (Accessed: 09 April 2024).
- Cimiano et al 2020a** Cimiano, Philipp, Chiarcos, Christian, McCrae, John P., and Gracia Jorge. (2020a) "Converting Language Resources into Linked Data" in *Linguistic Linked Data*, pp. 163–180. Cham: Springer International Publishing. Available at: https://doi.org/10.1007/978-3-030-30225-2_9 (Accessed: 09 April 2024).
- Cimiano et al 2020b** Cimiano, Philipp, Chiarcos, Christian, McCrae, John P., and Gracia, Jorge. (2020b) *Linguistic Linked Data: Representation, Generation and Applications*. Cham: Springer International Publishing. Available at: <https://doi.org/10.1007/978-3-030-30225-2> (Accessed: 09 April 2024).
- Costa 2013** Costa, Rute. (2013) "Terminology and Specialised Lexicography: two complementary domains". *Lexicographica* 29 (2013), pp. 29–42. De Gruyter. Available at: <https://doi.org/10.1515/lexi-2013-0004> (Accessed: 09 April 2024).
- Declerck et al 2020** Declerck, Thierry, McCrae, John Philip, Hartung, Matthias, Gracia, Jorge, Chiarcos, Christian, Montiel-Ponsoda, Elena, Cimiano, Philipp, Revenko, Artem, Saurí, Roser, Lee, Deirdre, Racioppa, Stefania, Abdul Nasir, Jamal, Orlikowsk, Matthias, Lanau-Coronas, Marta, Fäth, Christian, Rico, Mariano, Fazleh Elahi, Mohammad, Khvalchik, Maria, Gonzalez, Meritxell, and Cooney, Katharine. (2020) "Recent developments for the linguistic linked open data infrastructure" in *Proceedings of the Twelfth Language Resources and Evaluation Conference*, pp. 5660–5667, Marseille, France. European Language Resources Association. Available at: <https://aclanthology.org/2020.lrec-1.695.pdf> (Accessed: 09 April 2024).
- di Buono et all 2020** di Buono, Maria Pia, Cimiano, Philipp, Fazleh Elahi, Mohammad, and Grimm, Frank. (2020) "Terme-à-LLOD: Simplifying the Conversion and Hosting of Terminological Resources as Linked Data" in *Proceedings of the 7th Workshop on Linked Data in Linguistics (LDL-2020)*, Marseille, France, pp.28–35. European Language Resources Association. Available at: <https://aclanthology.org/2020.ldl-1.5.pdf> (Accessed: 09 April 2024).
- Di Nunzio and Vezzani 2021** Di Nunzio, Giorgio Maria and Vezzani, Federica. (2021) "One Size Fits All: A Conceptual Data Model for Any Approach to Terminology". *arXiv:2112.06562 [cs]*. Available at: <https://arxiv.org/pdf/2112.06562.pdf> (Accessed: 09 April 2024).
- Faber and L'Homme 2022** Faber, Pamela and L'Homme, Marie-Claude. (2022) *Theoretical Perspectives on Terminology*. John Benjamins Publishing Company.
- Frey and Hellmann 2021** Frey, Johannes and Hellmann, Sebastian. (2021) "FAIR Linked Data - Towards a Linked Data Backbone for Users and Machines" in *Companion Proceedings of the Web Conference 2021*. New York, NY, USA, pp. 431–435. Association for Computing Machinery. Available at: <https://doi.org/10.1145/3442442.3451364> (Accessed: 09 April 2024).
- Khan 2018** Khan, Anas Fahad. (2018) "Towards the Representation of Etymological Data on the Semantic Web", *Information*, 9(12), p. 304.
- Kitanović et al 2021** Kitanović, Olivera, Stanković, Ranka, Tomašević, Aleksandra, Škorić, Mihailo, Babić, Ivan, and Kolonja, Ljiljana. (2021) "A Data Driven Approach for Raw Material Terminology", *Applied Sciences*, 11 (7), 2892. Available at: <https://www.mdpi.com/2076-3417/11/7/2892> (Accessed: 09 April 2024).
- Montiel-Ponsoda et al 2011** Montiel-Ponsoda, Elena, Aguado De Cea, Guadalupe Gómez-Pérez, Asunción, and Peters, Wim. (2011) "Enriching ontologies with multilingual information", *Natural language engineering*, 17(3), pp. 283–309. Available at: <https://doi.org/10.1017/S1351324910000082> (Accessed: 09 April 2024).
- Piccini, Vezzani, and Bellandi 2023** Piccini, Silvia, Vezzani, Federica, and Bellandi, Andrea. (2023) "TBX and 'Lemon': What perspectives in terminology?" *Digital Scholarship in the Humanities*, 38 (Supplement_1), pp. i61–i72. Available at: <https://doi.org/10.1093/llc/fqad025> (Accessed: 09 April 2024).
- Rastier 1995** Rastier, Francois. (1995) "Le terme : entre ontologie et linguistique". *La banque de mots* 7, pp. 35–65.
- Reineke and Romary 2019** Reineke, Detlef and Romary, Laurent. (2019) "Bridging the gap between SKOS and TBX". *Die Fachzeitschrift für Terminologie* 19(2). Available at: <https://inria.hal.science/hal-02398820> (Accessed: 09 April 2024).
- Rodriguez-Doncel et al 2018** Rodriguez-Doncel, Víctor, Santos, Cristiana, Casanovas, Pompeu, Gómez-Pérez, Asunción, and Gracia, Jorge. (2018) "A Linked Data Terminology for Copyright Based on Ontolex-Lemon". in *AI Approaches to the Complexity of Legal Systems, Lecture Notes in Computer Science*, Cham, pp. 410–423. Springer International Publishing. Available at: https://doi.org/10.1007/978-3-030-00178-0_28 (Accessed: 09 April 2024).

- Santos and Costa 2015** Santos, Claudia and Costa, Rute. (2015) “Semasiological and onomasiological knowledge representation: Domain specificity”. in *Handbook of Terminology: Volume 1, Handbook of Terminology*, pp. 153–179. John Benjamins Publishing Company. Available at: <https://doi.org/10.1075/hot.1.dom1> (Accessed: 09 April 2024).
- Siemoneit, McCrae, and Cimiano 2015** Siemoneit, Benjamin, McCrae, John Philip, and Cimiano, Philipp. (2015) “Linking four heterogeneous language resources as linked data” in *Proceedings of the 4th Workshop on Linked Data in Linguistics: Resources and Applications*, pp. 59–63, Beijing, China. Association for Computational Linguistics. Available at: <https://aclanthology.org/W15-4207.pdf> (Accessed: 09 April 2024).
- Speranza et al 2020** Speranza, Giulia, Di Buono, Maria Pia, Sangati, Federico, and Monti, Johanna. (2020) “From linguistic resources to ontology-aware terminologies: Minding the representation gap” in *Proceedings of The 12th Language Resources and Evaluation Conference*, pp. 2503–2510. European Language Resources Association. Available at: <https://aclanthology.org/2020.lrec-1.305.pdf> (Accessed: 09 April 2024).
- Suchowolec, Lang, and Schneider 2019** Suchowolec, Karolina, Lang, Christian, and Schneider, Roman. (2019) “An empirically validated, onomasiologically structured, and linguistically motivated online terminology: Re-designing scientific resources on German grammar”. *International Journal on Digital Libraries*. 20(3), pp. 253–268. Available at: <https://doi.org/10.1007/s00799-018-0254-x> (Accessed: 09 April 2024).
- Vellutino et al 2016** Vellutino, Daniela, Maslias, Rodolfo, Rossi, Francesco, Mangiacapre, Carmina, and Montoro, Maria Pia. (2016) “Verso l'interoperabilità semantica di IATE. Studio preliminare sul lessico dei Fondi strutturali e d'Investimento Europei (Fondi SIE)”, *Diacronia XIII* (1), 187. Available at: <https://www.diacronia.ro/en/indexing/details/A23887/pdf> (Accessed: 09 April 2024).
- Wilkinson et al 2016** Wilkinson, Mark D., Michel Dumontier, IJsbrand Jan Aalbersberg, Gabrielle Appleton, Myles Axton, Arie Baak, Niklas Blomberg, et al (2016) “The FAIR Guiding Principles for Scientific Data Management and Stewardship”. *Scientific Data*, 3 (1): 160018. Available at: <https://doi.org/10.1038/sdata.2016.18> (Accessed: 09 April 2024).



This work is licensed under a Creative Commons Attribution-NonCommercial 4.0 International License.