LDK 2023

The Importance of Being Interoperable: Theoretical and Practical Implications in Converting TBX to OntoLex-Lemon

Andrea Bellandi
Cnr-Istituto di Linguistica Computazionale
"Antonio Zampolli", Italy
andrea.bellandi@ilc.cnr.it

Silvia Piccini
Cnr-Istituto di Linguistica Computazionale
"Antonio Zampolli", Italy
silvia.piccini@ilc.cnr.it

Giorgio Maria Di Nunzio
Dip. di Ingegneria dell'Informazione
Università di Padova, Italy
giorgiomaria.dinunzio@unipd.it

Federica Vezzani
Dip. di Studi Linguistici e Letterari
Università di Padova, Italy
federica.vezzani@unipd.it

Abstract

This paper introduces a methodology, design, and implementation of an interactive converter for transforming terminological data from the TermBase eXchange (TBX) format to the Ontolex-Lemon model. The paper highlights the differences between the two models, emphasizing their different technologies and data structures.

The proposed software architecture implements the conversion process through three main phases: analysis, filtering, and assembling. The analysis phase includes parsing the TBX file and generating an intermediate representation stored in a SQLite database. The filtering phase allows users to query and filter the data on the basis of their specific requirements. Finally, the assembling phase builds the Ontolex-Lemon lexicon by processing the filtered data and serializing it as RDF triples.

The converter aims to enable end users to actively participate in the conversion process, particularly in complex decision-making steps dealing with term variation, polysemy, and sense-concept relations.

1 Introduction

In the last decade Linked Data (LD) has been confirmed as one of the promising approaches for representing and connecting research data and metadata (Frey and Hellmann, 2021). In the context of linguistic resources, Linguistic Linked Open Data (LLOD) is a paradigm that promotes the publication and interlinking of resources such as e-lexicons, corpora, and terminologies. LLOD allows for a standardized way to access data, enabling researchers to explore, analyze, and utilize linguistic data for various language-related applications (Cimiano et al., 2020). Among the various data models, the Ontolex-Lemon model has gained popularity as the de-facto standard for representing lexical data using the Resource Description Framework (RDF) to express the information on the Semantic Web as LD (McCrae et al., 2017). However, there are specific cases where some types of linguistic resources have their own standard formats. This is the case of terminological resources encoded according to the TermBase eXchange (TBX) ISO standard 30042[1] – an XML-based family of terminology exchange formats compliant with the Terminological Markup Framework (TMF - ISO 16642: 2017)[2]. TBX, as well as other LD approaches, ensures consistency and interoperability by establishing a common structure and vocabulary for describing terminology across different systems and applications.

A number of methods and approaches, like for example the TBX2RDF conversion system (Cimiano et al., 2015; Montiel-Ponsoda et al., 2015), have been proposed to convert terminological data from the XML-based TermBase eXchange (TBX) format to Ontolex-Lemon, enabling their integration into the linguistic Linked Data ecosystem. Guidelines for a virtualization approach known as Term-à-LLOD have been developed to facilitate this conversion process (di Buono et al., 2020). In addition, there have been recent efforts to enhance Ontolex-Lemon with a dedicated module for representing terminology information[3].

Our proposal focuses on the mismatches between the two representations (one terminographical the other lexicographical) that, in order to be tackled and solved, require a necessary intervention of the user. In fact, these mismatches call into ques-

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

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

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

tion theoretical aspects that have been neglected by the previous works and that instead require active decisions by the scholars interested in converting their own data. In particular, the theoretical aspects related to this work have been discussed in a seminal paper (Piccini et al., 2023) and have been taken up, inspiring the preliminary design and implementation of such tool (Bellandi et al., 2023). We report here a brief summary of the considerations presented by (Bellandi et al., 2023):

- **lexicographical vs. terminological view**. A purely terminological vision (TBX) is transformed into a lexicographic standpoint (Ontolex-Lemon), where the conceptual dimension is not longer central and, conversely, sense acquires a crucial role.

- **ontology reuse**. The LD paradigm strongly encourages the reuse of existing vocabularies. According to this principle, the converter should make it possible to decide which data categories to use.

- **deductive rules**. The structure of the TBX file has some implicit relations among terms that get lost in the conversion from TBX to OntoLex-Lemon. The most important one is the information about synonymy among terms.

- **knowledge extraction**. In some cases the terminographer does not have a specific data category available in the TBX file to describe a particular behavior of the term. In such cases he/she can simply use the «note» field to store that information.

- **enriching the TBX**. After the knowledge extraction from unstructured notes, we can enrich the original TBX as well as its OntoLex-lemon counterpart with the new extracted information.

In this paper, we focus on the methodology, design, and implementation of the interactive converter that will allow terminologists to actively participate in the conversion process. In particular, we describe the conversion steps that require the user to make decisions about aspects such as variation, polysemy, and sense-concept relations.

2 How do TBX and *lemon* Differ

In this section, we briefly summarize the differences between TBX and OntoLex-lemon.

A basic key difference between the two models lies in their underlying technologies: TBX utilizes XML as its representation language, while Ontolex-Lemon is based on RDF and leverages the semantic capabilities of the Semantic Web. This distinction influences the way data is structured and the interoperability possibilities with other linked data resources. However, it is important to recognize that converting TBX to LD involves more than a shift from an XML-based to an RDF-based structure; it requires theoretical reflection and consideration of the conceptual and organizational differences between the two models (Piccini et al., 2022). In fact, the organizational differences are also reflected by the aim of the two models: TBX primarily emphasizes the exchange and management of terminological resources, ensuring consistency and interoperability among terminologists and language professionals. In contrast, Ontolex-Lemon is specifically tailored for representing lexical data, aiming to capture detailed linguistic information and to enable semantic integration with other RDF datasets.

The objective of this paper is therefore to examine the prerequisites of a converter capable of processing the latest editions of TBX and Ontolex-Lemon. The analysis will particularly concentrate on the theoretical consequences that arise from the shift from a structure centered on concepts (TBX) to one centered on senses (Ontolex-Lemon).

3 Towards a TBX to *lemon* Converter

Given the different nature of the two models, we propose to create an interactive and configurable converter that can indulge the theoretical vision of the user who carry out the conversion, whether they are terminologists, translators, or lexicographers. In light of this, converting a TBX resource to Ontolex-Lemon should require a dedicated software architecture as depicted in Figure 1. The latter translates a TBX source into RDF triples, going interactively through three main phases: i) *analysis*, ii) *filtering*, and iii) *assembling*.

3.1 Phase 1: Analysis

Concerning the first phase, the parser component is in charge of analyzing the XML input file, potentially written in different TBX public dialects

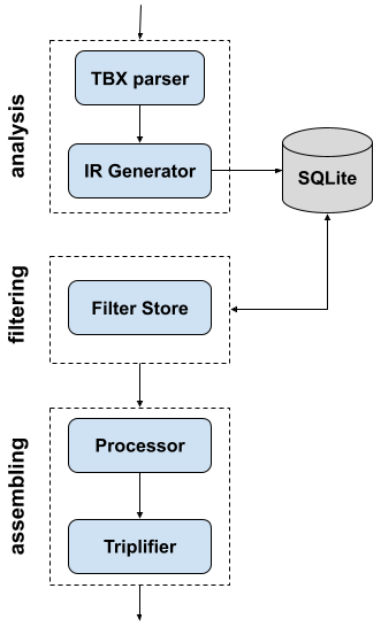

Figure 1: The architecture of the three phases of the converter from a TBX to an Ontolex-lemon representation.

(core, min, basic), and is aimed at producing an intermediate representation (IR) of the information contained. IR represents a partial conversion of the TBX elements such as concepts, terms, and languages, in a series of RDF triples, without making any assumption on the final output.

3.2 Phase 2: Filtering

IR is stored in a SQLite database, together with some metadata (for example transaction types, creation dates, subject fields), allowing the *filtering* phase to implement fast and feasible querying for user-specific filters to select and eventually enrich the data.

3.3 Phase 3: Assembling

Starting from the filtered data, the third phase constructs the Ontolex-Lemon lexicon by processing the languages, the concepts, and the terms (the Processor component in Figure 1), and serializes them as RDF triples according to the Ontolex-Lemon data model (the triplifier component in Figure 1).

The Processor is the crucial component of the software architecture because it is in charge of taking into account the desiderata of the user who makes the conversion. It potentially can be composed of a pipeline of processors that implements those desiderata starting from the IR data, for example:

- bypassing the Ontolex-Lemon Lexical Sense class and linking lexical entries directly to the designated concepts,
- linking the terms denoting the same concept across different languages by means of the translation property,
- creating polysemous entries in Ontolex-Lemon in those cases in which the terms designate different concepts but are characterized by the same orthographic form and share the same etymology,
- creating relationships of synonymy between terms designating the same concept in a given language.

Currently, the software prototype performs a conversion process based on the default behavior. The following section is devoted to presenting a simple example of default conversion.

4 A Conversion Example

The hierarchical structure of a TBX file is basically the following:

- a set of concept entries (tag <conceptEntry>),
- within each concept entry, a set of language sections (tag <langSec>),
- for each language section, a set of terms that designate the concept for that language (tag <termSec>).

Figure 2 depicts a fragment of an example of a TBX-basic terminological database with one concept. In particular,

- the fragment, reports a concept called *c1*, related to the e-mobility field,
- and two language sections, for English and French, with their respective definitions for that concept.
- There are two terms for concept *c1* in English, neighborhood "car vehicle" and "NEV", while one in French, "véhicule de proximité". For each term, some kind of information is specified, such as morphology, term type, and administrative status.

```
<conceptEntry id="c1">
    <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>
        </descripGrp>
        <termSec>
            <term>neighborhood electric vehicle</term>
            <basic:termType>fullForm</basic:termType>
            <min:partOfSpeech>noun</min:partOfSpeech>
            <basic:grammaticalGender>masculine
            </basic:grammaticalGender>
            <min:administrativeStatus>preferredTerm-admn-sts
            </min:administrativeStatus>
        </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">
        <descripGrp>
            <basic:definition>Véhicule à deux places,
            activé par un moteur électrique à courant
            continu alimenté par des batteries au plomb
            rechargeables à partir d'une prise de courant
            résidentielle de 110 volts.
        </basic:definition>
        <termSec>
            <term>véhicule de proximité</term>
            <basic:termType>fullForm</basic:termType>
            <min:partOfSpeech>noun</min:partOfSpeech>
            <min:administrativeStatus>admittedTerm-admn-sts
            </min:administrativeStatus>
        </termSec>
    </langSec>
</conceptEntry>
```

Figure 2: A TBX-basic dialect example.

Our converter performs the conversion and the result is reported in Figure 3. RDF triples are encoded in turtle syntax, and they are grouped according to the TBX entities they correspond to.

Concerning the <conceptEntry> entity, concepts are converted by means of the SKOS ontology, according to Reineke and Romary (2019). All the subject fields correspond to SKOS concept schemes, while concepts are mapped to SKOS concepts. The membership of concepts to their subject fields is formalized through the SKOS *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. Figure 2 reports an example related to the latter case. A definition of the concept in each language is formalized as Figure 3 shows. Other TBX data categories, such as note, source, and cross reference, are mapped to SKOS *note*, Dublin core *source*, and RDF *seeAlso* properties, respectively.

Concerning the <langSec> entity, the related *lemon* lexica are created. Referring to the example in Figure 2, both English and French lexica are defined as in the second group of triples in Figure 3. Furthermore, the terms of each language are defined as entries of the suitable lexicon. If the definition contained in the <langSec> had had a source or/and an external reference, we would have used the reification mechanism[4] in order to represent the source and the reference of the concept definition, by means of Dublin core *source*, and RDF *seeAlso* properties, respectively.

concepts

```
:c1 a skos:Concept ;
   skos:prefLabel "c1"@en ;
   skos:inScheme :sbjf_1 ;
   skos:definition "A battery .."@en ;
   skos:definition "Véichule à deux .."@fr .

:sbjf_1 a skos:ConceptScheme ;
   skos:prefLabel "e-mobility"@en .
```

languages

```
:lexEN a lime:Lexicon ;
   dct:language "en" ;
   lime:entry :t1, :t2 .

:lexFR a lime:Lexicon ;
   dct:language "fr" ;
   lime:entry :t3 .
```

terms

```
:t1 a ontolex:LexicalEntry ;
   lexinfo:partOfSpeech lexinfo:noun ;
   lexinfo:termType lexinfo:fullForm ;
   ontolex:canonicalForm [
            lexinfo:gender lexinfo:masculine ;
      ontolex:writtenRep "neighborhood
                  electric veichle"@en .
   ] ;
   ontolex:sense :t1_sense .

:t1_sense a ontolex:LexicalSense ;
            lexinfo:normativeAuthorization
              lexinfo:preferredTerm .

:t2 a ontolex:LexicalEntry ;
   lexinfo:partOfSpeech lexinfo:noun ;
   lexinfo:termType lexinfo:acronym ;
   ontolex:canonicalForm [
      ontolex:writtenRep "NEV"@en .
   ] ;
   ontolex:sense :t2_sense .

:t2_sense a ontolex:LexicalSense ;
            lexinfo:normativeAuthorization
              lexinfo:admittedTerm .

:t2 lexinfo:acronymFor :t1 .

t3 a ontolex:LexicalEntry ;
   lexinfo:partOfSpeech lexinfo:noun ;
   ontolex:canonicalForm [
      ontolex:writtenRep "véichule de proximité"@fr .
   ] ;
   ontolex:sense :t3_sense .

:t3_sense a ontolex:LexicalSense ;
            lexinfo:normativeAuthorization
              lexinfo:admittedTerm .
```

Figure 3: The converted data in *lemon* is serialized by means of the turtle syntax.

Finally, terms contained in the <termSec> entity are represented as lexical entries in the Ontolex-Lemon model. Each term is mapped to a Lexical Entry element, without specifying its particular type (word or multi-word), and it is represented as a canonical form of that lexical entry. According to the "*semantics by reference*" paradigm of Ontolex-Lemon, the meaning of a lexical entry is

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

specified by referring to the created SKOS concept that represents its meaning. The default conversion process creates a lexical sense for each lexical entry and links it to the suitable concept by means of the *reference* property. Since the model does not contain a complete collection of linguistic categories, it relies on Lexinfo vocabulary[5]. As a consequence, morphological information, such as part of speech, gender, and number is associated with the forms, while usage context, term type, and administrative status are associated with the senses, according to the Lexinfo schema.

5 Conclusion and Future Work

In this paper, we have presented the current work on the definition of a methodology for the conversion of terminological data as well as the design and implementation of an interactive converter from TBX to Ontolex-Lemon. Despite the already available tools for this type of conversion, we believe that transforming TBX data to Ontolex-Lemon can be more challenging than just carefully mapping and transforming all of the (meta)data of the different elements from one model to another. In fact, the two different frameworks (TBX concept-oriented and Ontolex-Lemon sense-centered) necessitate a deep understanding of both models and the ability to reconcile the differences in their structures and semantics during the conversion process.

The current prototype of the conversion tool allows the user to explore and analyze the structure (what data categories are available) and the statistics (how many concepts, languages, and terms) of the TBX file. In addition, the user can also make some choices about the mapping and identification of TBX concepts into SKOS concepts across different languages and from TBX terms to Ontolex-lemon lexical concepts. As future work, we are currently working on parameterizing the default behavior on some steps such as:

- make explicit the choice of the use of Ontolex-lemon senses (or not);
- make explicit the decision of the management of synonymy and the equivalents across multiple languages;
- extrapolate information from TBX textual data categories (for example the element <note>) that can be mapped into Ontolex-lemon properties.

[5]LexInfo is an ontology that provides data categories for the *lemon* model. Please, see https://lexinfo.net/

6 Acknowledgment

This work has been carried out in the framework of agreement between Consiglio Nazionale delle Ricerche – Istituto di Linguistica Computazionale and RUT Foundation. This work is also part of the initiatives carried out by the Center for Studies in Computational Terminology (CENTRICO) of the University of Padua and in the research directions of the Italian Common Language Resources and Technology Infrastructure CLARIN-IT.

References

Andrea Bellandi, Giorgio Maria Di Nunzio, Silvia Piccini, and Federica Vezzani. 2023. 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.

Philipp Cimiano, Christian Chiarcos, John P. McCrae, and Jorge Gracia. 2020. *Linguistic Linked Open Data Cloud*, pages 29–41. Springer International Publishing, Cham.

Philipp Cimiano, John P. McCrae, Víctor Rodríguez-Doncel, Tatiana Gornostay, Asunción Gómez-Pérez, Benjamin Siemoneit, and Andis Lagzdins. 2015. Linked terminologies: applying linked data principles to terminological resources. In *Proceedings of the eLex 2015 Conference*, pages 504–517.

Maria Pia di Buono, Philipp Cimiano, Mohammad Fazleh Elahi, and Frank Grimm. 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)*, pages 28–35, Marseille, France. European Language Resources Association.

Johannes Frey and Sebastian Hellmann. 2021. FAIR Linked Data - Towards a Linked Data Backbone for Users and Machines. In *Companion Proceedings of the Web Conference 2021*, WWW '21, pages 431–435, New York, NY, USA. Association for Computing Machinery.

John McCrae, Julia Bosque-Gil, Jorge Gracia, Paul Buitelaar, and Philipp Cimiano. 2017. The OntoLex-Lemon Model: Development and Applications. In *Electronic lexicography in the 21st century: Lexicography from scratch. Proceedings of eLex 2017*, pages 587–597, Brno. Lexical Computing CZ s.r.o. Http://en.wikipedia.org/wiki/Galway; https://en.wikipedia.org/wiki/Leiden.

Elena Montiel-Ponsoda, Julia Bosque-Gil, Jorge Gracia, Guadalupe Aguado de Cea, and Daniel Vila-Suero. 2015. Towards the Integration of Multilingual Terminologies: an Example of a Linked Data Prototype. In *Terminology and Artificial Intelligence (TAI)*, pages 205–206.

Silvia Piccini, Federica Vezzani, and Andrea Bellandi. 2022. Entre TBX et Ontolex-Lemon : Quelles Nouvelles Perspectives en Terminologie? (poster). In *Proceedings of the 1st International Conference on Multilingual Digital Terminology Today*, volume 3161 of *CEUR Workshop Proceedings*, Padua, Italy. CEUR. ISSN: 1613-0073.

Silvia Piccini, Federica Vezzani, and Andrea Bellandi. 2023. TBX and 'Lemon': What perspectives in terminology? *Digital Scholarship in the Humanities*, 38(Supplement_1):i61–i72.

Detlef Reineke and Laurent Romary. 2019. Bridging the gap between SKOS and TBX. *edition - Die Fachzeitschrift für Terminologie*, 19(2). Publisher: Deutscher Terminologie-Tag e.V. (DTT).