Representing Multilingual Terminologies with OntoLex-Lemon

Patricia Martín-Chozas[1], Thierry Declerck[2]

[1]Ontology Engineering Group, Universidad Politécnica de Madrid, Avda. Monteprínicipe, s/n, Boadilla del Monte, 28660, Spain

[2]German Research Center for Artificial Intelligence GmbH (DFKI), Multilinguality and Language Technology Lab, Saarland Informatics Campus D3 2, Stuhlsatzenhausweg 3, 66123 Saarbrücken, Germany

Abstract

This paper is framed within a project to make multilingual terminologies available in a native graph representation format. We are exploring the use of the OntoLex-Lemon model, suggesting also some extensions, for achieving a declarative encoding of relations between multilingual expressions contained in terminologies. This model is not only used for encoding terms but also for their associated definitions, contexts and notes. With this effort, we aim at supporting the publication of multilingual terminologies in the Linked Open Data cloud.

Keywords

Terminologies, Multilingualism, Formal Representation, OntoLex-Lemon

1. Introduction

In the context of work dealing with the conversion of multilingual terminologies onto an RDF[1] model, we came into modelling decisions concerning also additional language data included in such resources. While the original purpose of the porting exercise is not to change anything at the level of the content of the considered terminologies, their modelling in a graph-based representation offers possibilities for their interlinking and merging with other resources, being in the realm of terminologies or of other types of data, like for example detailed lexicographic resources. Thus, the focus of our work is the possible improved formal representation of the language data used in multilingual terminologies. We discuss in this short paper few decisions points concerning our modelling strategy, also comparing our work with a directly related former approach.

1st International Conference on "Multilingual digital terminology today. Design, representation formats and management systems", June 16 – 17, Padova, Italy

pmchozas@fi.upm.es (P. Martín-Chozas); declerck@dfki.de (T. Declerck)
https://github.com/pmchozas/ (P. Martín-Chozas); https://www.dfki.de/~declerck/ (T. Declerck)
000-0002-8922-7521 (P. Martín-Chozas); 0000-0002-9450-6648 (T. Declerck)

[1] RDF stands for "Resource Description Framework". See https://www.w3.org/TR/rdf-primer/for more details.

2. The Data Basis: Two Terminological Resources

Currently, we consider two terminological resources as the input for our transformation work: the multilingual terminology of the Deutsche Bahn (German Railways), which is encoded within the TBX[2] standard and can be accessed online[3]; and IATE (**I**nter**a**ctive **T**erminology for **E**urope)[4], one of the most representative terminological database in Europe. The consideration of the latter was motivated by a previous exercise that focused on the conversion of the data contained in IATE, structured in TBX, into RDF. This effort is a great starting point to compare our approach.

3. The TBX2RDF Guidelines

The past LIDER project[5] was already concerned with mapping TBX to RDF, with the goal of transforming and publishing terminologies as Linked Data [2]. LIDER developed guidelines for this task[6] in which TBX elements are converted into OWL[7] and associated with other RDF vocabularies, while the basic vocabularies chosen as the backbone of the conversion were SKOS[8] and the *lemon* model [3], a predecessor of the OntoLex-Lemon framework [4] we are using. [5] describe the TBX2RDF approach[9] and [6] presents recent developments related to this initiative, relying on a virtualization approach that is making use of containerization technologies.

The LIDER TBX2RDF approach is representing the TBX terminological concepts as skos:Concept and the TIG/NTIG elements of TBX as ontolex:LexicalEntry, and most of the other TBX elements are straightforwardly mapped onto RDF, meaning that they are encoding as URIs for representing a resource that can be associated with RDF predicates and objects. We note also that TXB2RDF is not representing the TBX langSet data as such, but instead is creating language specific lexicons in which all the data included in the original langSet element are encoded.

4. Our Approach

We make use of the most recent version of OntoLex-Lemon,[10] which is effectively integrating the SKOS vocabulary for representing conceptual units and their associated language data. This was not the case with its former version, *lemon*, which was used in the LIDER project. We can now use properties defined in OntoLex-Lemon for directly linking the conceptually oriented

[2] TBX stands for "TermBase eXchange". See https://www.tbxinfo.net/ [accessed 2022-02-14], or [1] for more details.

[3] www.deutschebahn.com/dblanguageportal [accessed 2021-10-02]

[4] See https://iate.europa.eu/ [accessed 2022-02-14]

[5] http://lider-project.eu/lider-project.eu/index.html [accessed 2021-10-02]

[6] The latest version of those guidelines is available at https://github.com/bpmlod/report/blob/gh-pages/multilingual-terminologies/index.html [accessed 2022-02-14]

[7] OWL stands for "Web Ontology Language". See https://www.w3.org/TR/owl2-primer/ [accessed 2022-14-02]

[8] SKOS stands for „Simple Knowledge Organization System". See also https://www.w3.org/2009/08/skos-reference/skos.html [last consulted: 2022-02-14]

[9] The corresponding W3C Community Group Report is avaialable at https://www.w3.org/2015/09/bpmlod-reports/multilingual-terminologies/[accessed2022-02-14]

[10] See https://www.w3.org/2016/05/ontolex/ [accessed 2022-02-14] for technical details.

terms to lexical entries, while the LIDER TBX2RDF converter was using a custom property for this purpose. We introduce a skos:ConceptScheme for encoding the whole conceptual organisation of the original terminology, and within this scheme we allow for the definition of specific domain subsets, a feature not supported in TBX.[11] OntoLex-Lemon is foreseeing as a subclass of skos:Concept the class ontolex:LexicalConcept for linking lexical entries to the conceptual part described in the SKOS vocabulary. We encode all the terms as instances of this class, and no longer as instances of the class ontolex:LexicalEntry, as it was implemented in TBX2RDF. Another, and more significant, departure from the LIDER TBX2RDF model is the fact that we model definitions and contexts as instances of classes, and no longer as literal values. In doing so, we can describe specific relations between the definitions within one language or across different languages. In the latter case, we can specify if the definitions given for terms in two different languages are translations of each other, multilingual equivalents or just monolingual definitions included in the multilingual terminology. Suggested additions to the OntoLex-Lemon model are marked with the prefix "termlex".

Figure 1 shows how an IATE term entry is currently represented following our approach, while also representing the synonymy of two Spanish terms. Figure 2 displays the relations between the terms and their definitions, which as instances of a class, can link to further information, like the provenance or the definitions for the same original term entry in another language. The English equivalents for the Spanish terms "surco ferroviario" and "franja ferroviaria" (displayed in Figures 1 and 2) – "train path", "train slot" –, as well as the English definitions and their context of use are linked to the Spanish terms and entries via the properties defined in the Vartrans module of OntoLex-Lemon,[12] supporting a declarative description of the different types of relations that can exist between those different types of language data (terms, definitions and contexts of use).

5. Conclusions and Future Work

We described ongoing work in porting the multilingual terminology resources onto a Linked Data compliant representation language. This work led us to the question if it would not be suitable to extend the modelling of TBX terminologies in RDF already proposed by the LIDER TBX2RDF converter. One aspect consists in considering definitions, contexts and notes as full ontological elements that can thus be put explicitly in relation to each other. This way, definitions in different languages can be declaratively interlinked and marked as translations, equivalents or as not having any of those relations.

As an outcome of our work, we are currently proposing an extension module for OntoLex-Lemon,[13] that deals with the representation of terminological data that is not covered in the core module, as the main motivation of the development of OntoLex-Lemon vocabulary was to represent language data with references to ontologies.

[11] See [7] for a discussion on the difference between the "subjectField" in TBX and the conceptual hierarchy in SKOS.

[12] https://www.w3.org/2016/05/ontolex/#variation-translation-vartrans

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

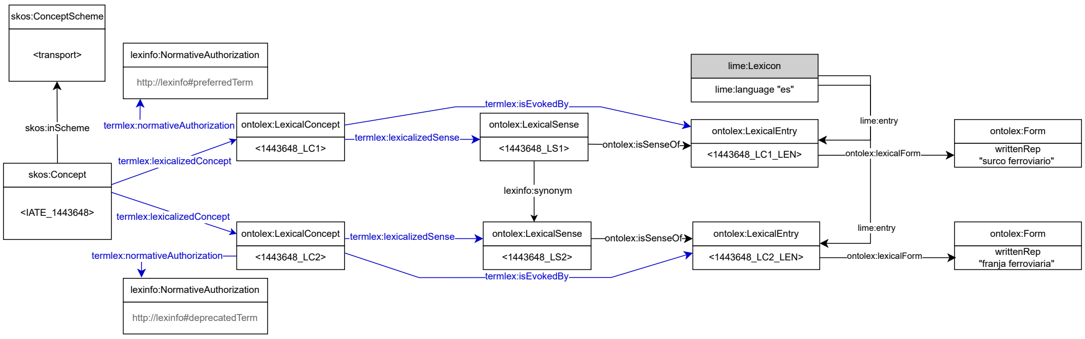

Figure 1: Representing a IATE term entry in OntoLex-Lemon, showing two Spanish terms used for a term entry. One term is marked as "preferred" while the other is marked as "deprecated". Our suggested extensions ("termlex") to OntoLex-Lemon are displayed in blue colour.

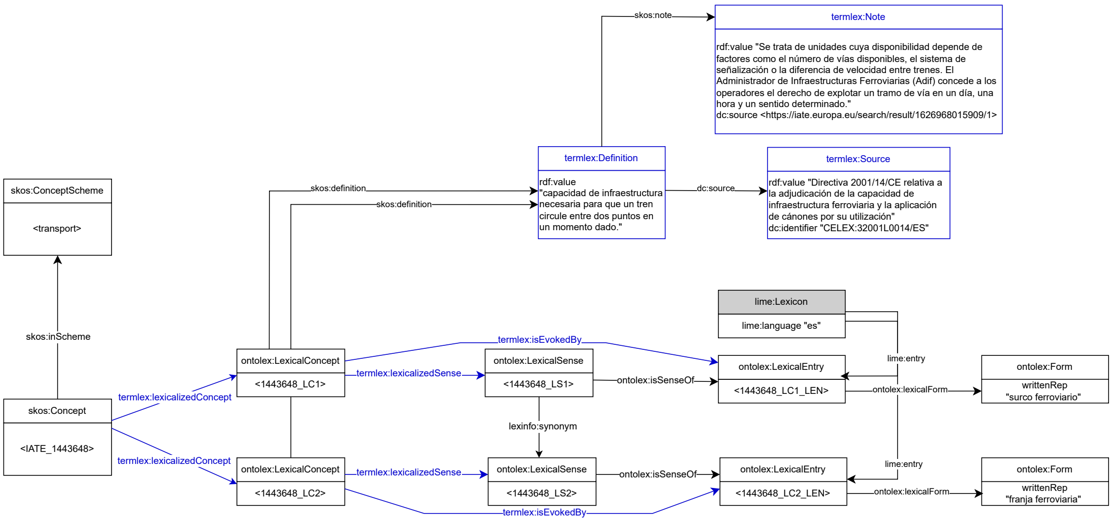
skos:note
termlex:Note
rdf:value "Se trata de unidades cuya disponibilidad depende de factores como el número de vías disponibles, el sistema de señalización o la diferencia de velocidad entre trenes. El Administrador de Infraestructuras Ferroviarias (Adif) concede a los operadores el derecho de explotar un tramo de vía en un día, una hora y un sentido determinado."
dc:source <https://iate.europa.eu/search/result/1626968015909/1>
termlex:Definition
rdf:value "capacidad de infraestructura necesaria para que un tren circule entre dos puntos en un momento dado."
dc:source
termlex:Source
rdf:value "Directiva 2001/14/CE relativa a la adjudicación de la capacidad de infraestructura ferroviaria y la aplicación de cánones por su utilización"
dc:identifier "CELEX:32001L0014/ES"
skos:definition
skos:definition
skos:ConceptScheme
<transport>
skos:inScheme
skos:Concept
<IATE_1443648>
termlex:lexicalizedConcept
termlex:lexicalizedConcept
ontolex:LexicalConcept
<1443648_LC1>
ontolex:LexicalConcept
<1443648_LC2>
termlex:isEvokedBy
termlex:lexicalizedSense
ontolex:LexicalSense
<1443648_LS1>
lexinfo:synonym
termlex:lexicalizedSense
ontolex:LexicalSense
<1443648_LS2>
termlex:isEvokedBy
ontolex:isSenseOf
ontolex:isSenseOf
lime:Lexicon
lime:language "es"
lime:entry
lime:entry
ontolex:LexicalEntry
<1443648_LC1_LEN>
ontolex:LexicalEntry
<1443648_LC2_LEN>
ontolex:lexicalForm
ontolex:lexicalForm
ontolex:Form
writtenRep
"surco ferroviario"
ontolex:Form
writtenRep
"franja ferroviaria"

Figure 2: Representing the links between terms and their definitions, which are now instances of a specific class. Our suggested extensions ("termlex") to OntoLex-Lemon are displayed in blue colour.

Acknowledgments

This short paper is based upon work from COST Action NexusLinguarum – European network for Web-centered linguistic data science (CA18209), supported by COST (European Cooperation in Science and Technology). The article is also supported by the Horizon 2020 research and innovation programme with the project Prêt-à-LLOD (grant agreement no. 825182).

References

[1] A. Lommel, A. K. Melby, N. Glenn, J. Hayes, T. Snow, TBX-Min: A Simplified TBX-Based Approach to Representing Bilingual Glossaries, in: Terminology and Knowledge Engineering 2014, Berlin, Germany, 2014, p. 10 p. URL: https://hal.archives-ouvertes.fr/hal-01005851.

[2] C. Bizer, T. Heath, T. Berners-Lee, Linked data: The story so far, in: Semantic services, interoperability and web applications: emerging concepts, IGI global, 2011, pp. 205–227.

[3] J. P. McCrae, G. Aguado de Cea, P. Buitelaar, P. Cimiano, T. Declerck, A. Gómez-Pérez, J. Gracia, L. Hollink, E. Montiel-Ponsoda, D. Spohr, T. Wunner, Interchanging lexical resources on the semantic web, Lang. Resour. Evaluation 46 (2012) 701–719. URL: https://doi.org/10.1007/s10579-012-9182-3. doi:10.1007/s10579-012-9182-3.

[4] J. P. McCrae, P. Buitelaar, P. Cimiano, The OntoLex-Lemon Model: development and applications, in: Proc. of the 5th Biennial Conference on Electronic Lexicography (eLex), 2017.

[5] P. Cimiano, J. P. McCrae, V. Rodríguez-Doncel, T. Gornostay, A. Gómez-Pérez, B. Siemoneit, A. Lagzdins, Linked terminologies: applying linked data principles to terminological resources, in: Proceedings of the eLex 2015 Conference, 2015.

[6] M. P. di Buono, P. Cimiano, M. F. Elahi, F. Grimm, 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), European Language Resources Association, Marseille, France, 2020, pp. 28–35. URL: https://aclanthology.org/2020.ldl-1.5.

[7] D. Reineke, L. Romary, Bridging the gap between SKOS and TBX, edition - Die Fachzeitschrift für Terminologie 19 (2019). URL: https://hal.inria.fr/hal-02398820.