Linked Terminologies: Applying Linked Data Principles to Terminological Resources

Philipp Cimiano[1], John P. McCrae[1,4], Víctor Rodríguez-Doncel[2], Tatiana Gornostay[3], Asunción Gómez-Pérez[2], Benjamin Siemoneit[1], Andis Lagzdins[3]

[1]Cognitive Interaction Technology, Excellence Cluster, Bielefeld University, Germany
[2]Ontology Engineering Group, Universidad Politecnica de Madrid, Spain
[3]Tilde, Latvia
[4]National University of Ireland, Galway, Ireland
{cimiano,jmccrae}@cit-ec.uni-bielefeld.de, bsiemone@techfak.uni-bielefeld.de,
{vrodriguez,asun}@fi.upm.es, {tatjana.gornostaja,andis.lagzdins}@tilde.lv

Abstract

In this paper we present an approach to publishing and linking terminological resources using linked data principles. We describe how terminologies can be represented in the Resource Description Framework (RDF), and as proof-of-concept we describe the application of these principles to two well-known terminologies, that is the InterActive Terminology for Europe (IATE) and the European Migration Network (EMN) glossary. We further present a simple yet effective method for inducing links between terminologies and present a small evaluation of the quality of the automatically induced links. We also present a publicly available service to transform TBX documents into RDF that we have used for the conversion of IATE to RDF.

Keywords: terminology; linked data; TBX; IATE; EMN

1. Introduction

Terminological resources (*terminologies* further in the text) play an important role in many applications where terminological consistency needs to be achieved or content needs to be described in multiple languages, for different audiences, levels of expertise, etc. So far, however, it is not trivial to discover, combine and exploit multiple terminologies within one application, nor is it easy to bootstrap the creation or extension of existing terminologies with content from other terminologies. To support such scenarios, an important step is to ensure that terminologies do not exist independently of each other, but are mutually linked to form a larger ecosystem of many (linked) terminologies comprising many domains, languages, etc.

Providing a first step towards creating such an ecosystem of linked terminologies, in this paper we propose a novel approach to publish and manage terminological datasets as linked data. Linked data represents a new paradigm for publishing data on the web relying on Semantic Web standards (RDF[1] and SPARQL[2]) in such a way that data is linked across

[1] http://www.w3.org/RDF/

[2] SPARQL is the query language for the RDF data model, see http://www.w3.org/TR/rdf-sparql-query/

datasets and sites. The main principles of Linked Data as defined by Tim Berners-Lee, the inventor of the World Wide Web, are as follows[3] (Heath and Bizer, 2011):

1. Entities in the data should be named via unique URIs;
2. These URIs should be HTTP URIs and resolve using standard web protocols;
3. When these URIs are resolved, they should return useful information about the resource;
4. They should contain links to other URIs so people can discover related resources.

We apply linked data principles to terminological datasets and present an approach to transform term bases in TBX format to RDF. Our approach is based on the *lemon* model[4] (McCrae et al., 2011), an RDF model developed to support publishing lexical resources as linked data. The proposed methodology has been implemented as an online service named TBX2RDF. We provide proof-of-concept for this transformation using the well-known InterActive Terminology for Europe (IATE) term base as well as the European Migration Network (EMN) glossary. While IATE was already available in TBX format, the EMN glossary was not, and it was directly converted from HTML into RDF format. The Linked Data version of IATE is available at `http://tbx2rdf.lider-project.eu/data/iate`, and the Linked Data version of the EMN glossary is also available online: `http://data.lider-project.eu/emn`. An implementation of the four linked data principles mentioned above can be exemplified with the URI `http://tbx2rdf.lider-project.eu/data/iate/competence+of+the+Member+States-en`, it uniquely identifies the lexical entry *'Competence of the Member States'* within IATE, it is resolvable, and the returned message provides information on the resource, being additionally linked to other URIs.

We also present an automatic method to link different terminological datasets to each other. This contributes to the creation of a seamless ecosystem of terminologies that can be easily accessed and navigated and creates added value by allowing applications to access and exploit a network of linked terminologies. To show the advantages of this linking, we include the links directly into the Linked Data version of IATE as well as the EMN dataset, so that users exploring one of these can navigate to related terms of the other resource. By linking also to the Manually Annotated Subcorpus (MASC) of the American National Corpus (ANC), we also show that our approach can be extended to linking terminologies to the mentions of the terms in a corpus.

It is important to mention that we are not proposing to replace TBX by a new format. In fact, we regard our work as providing an alternative serialization of terminologies in RDF format. We assume that terminologies will be natively stored and managed using the TBX data model, but that in addition they will be exposed in RDF to support the linking of terminologies across datasets, thus supporting the creation of the above mentioned ecosystem.

[3] `http://www.w3.org/DesignIssues/LinkedData.html`

[4] `http://lemon-model.net/`

When we started this project, we were surprised to see that there was no standard and agreed-upon format for publishing terminologies as RDF. One possibility would have been to develop an RDF model that is faithful to the original TBX model, reusing essentially the data schema behind TBX. However, this would have reduced interoperability with other lexical resources published as Linked Data including bilingual dictionaries, monolingual dictionaries, wordnets, etc. To support this, we have reused existing vocabularies for representing lexical information in connection to ontologies (e.g. the lexicon model for ontologies or lemon for short) as well as vocabularies to describe provenance of data and transaction information (e.g. the PROV-O ontology).

In essence, the main advantage we see in publishing terminologies as RDF is that this supports linking across datasets. While one might argue that the links in some sense are already *'hidden'* in the data as they are induced automatically on the basis of information available in the data in our approach, these links are made explicit as a result of this, so that others can directly exploit these links instead of having to recompute them. Further, in case links are provided by a third party between for example TBX and IATE, to where would these links be added? The third party might not have the right to add these links to the original dataset, so the links themselves would then have to be published as Linked data, clearly creating an added value that was not previously there.

In addition, RDF represents a very flexible data model that supports the flexible organisation of terminologies as a (directed) graph, allowing direct representation of terminological relations (such as *broader term*, *narrower term*, etc.) as edges in the RDF model. Second, using RDF as a data model eases the manipulation and handling of terminological data as standard tasks in terminology management can be broken down to SPARQL queries, such as: i) selecting the term entries in a particular language, ii) selecting corresponding terms in two given languages, iii) selecting the subset of a term base for a given subject field, iv) finding duplicate term entries, or v) selecting all deprecated terms in a particular resource. Further, moving to a datamodel such as RDF offers additional flexibility in that copyright and licensing information can be specified at the level of each term and term entry (Cabrio et al., 2014; Rodriguez-Doncel et al., 2014), allowing to include terms with different status and provenance within one resource, thus supporting fine-grained specification of provenance and licensing information.

The paper is structured as follows: we describe our proposed model for representing terminologies in RDF in Section 2. We then discuss in Section 3 how two terminologies have been migrated into RDF based on the lemon model as proof-of-concept. Section 4 describes our methodology for linking the terminologies to each other as well as to BabelNet and MASC, and includes a small evaluation in terms of precision of the induced links. We present a publicly available service for transforming terminologies in TBX format into RDF in Section 5, concluding in Section 6.

2. Representation of terminologies in RDF

In this section, we describe how terminologies can be represented using the Resource Description Framework (RDF). For the sake of presentation, we assume that terminologies are given in the TBX format, which is an open XML format for terminologies originally specified by the now defunct Localization Industry Standards Association (LISA)[5], and now available as an ISO standard (ISO, 2008). This does not represent any restriction as other formats can be converted to the proposed representation. This is corroborated by the fact that the European Migration Network terminology that we consider in Section 3 was not natively available in TBX, but only via HTML, which we transformed into lemon/RDF.

Our proposed representation for terminologies in RDF, fully described online[6], relies on the *lemon* vocabulary. *Lemon* stands for the *Lexicon Model for Ontologies* (McCrae et al., 2011) and was designed to represent lexical information in combination with ontologies. *lemon* meets the needs for representing terminologies in RDF as the conceptual backbone of a terminology can be regarded as an ontology. The terms themselves can be regarded as lexical elements, and are represented in *lemon* as *lexical entries*.

In what follows, we describe the representation of terminologies in RDF in a step-by-step fashion. For the purpose of this section we will discuss the conversion to RDF using the sample terminology in TBX format in Figure 1. We start by describing how terminological concepts are represented in our RDF representation.

The term entry in lines 3–7 would be represented in RDF by a `skos:Concept`. The Simple Knowledge Organization System (SKOS) is a vocabulary for representing knowledge organization systems (KOS) such as thesauri, classification schemes, subject heading and taxonomies in RDF. The fundamental element of a SKOS vocabulary are *concepts*, defined as *'units of thought, ideas, meanings, or (categories of) objects and events, which underlie many knowledge organization systems'*. As terminologies can be seen as a special case of a knowledge organization system, using SKOS concepts to represent terminological concepts seems appropriate.

This is shown by the following RDF snippet, where the the subject field of the terminological concept is specified via the property `subjectField`:

```
:IATE_84
  a  skos:Concept ;
  tbx:subjectField  "1011"^^xsd:string .
```

Our TBX document as shown in Figure 1 has two language sets for English and German. In the *lemon* model, a lexicon is regarded as language-specific and as comprising lexical entries

[5] http://www.ttt.org/oscarStandards/tbx/

[6] http://www.w3.org/community/bpmlod/wiki/Converting_TBX_to_RDF

```xml
<text>
  <body>
    <termEntry id="IATE-84">
      <descripGrp>
        <descrip type="subjectField">1011</descrip>
      </descripGrp>
    </termEntry>
    <langSet xml:lang="en">
      <tig>
        <term>competence of the Member States</term>
        <termNote type="termType">fullForm</termNote>
        <descrip type="reliabilityCode">3</descrip>
      </tig>
    </langSet>
    <langSet xml:lang="de">
      <ntig>
        <termGrp>
          <term>Zuständigkeit der Mitgliedstaaten</term>
          <termNote type="termType">fullForm</termNote>
          <descrip type="reliabilityCode">3</descrip>
          <termCompList type="lemma">
            <termCompGrp>
              <termComp>Zuständigkeit</termComp>
              <termNote type="partOfSpeech">noun</termNote>
              <termNote type="grammaticalNumber">singular</termNote>
            </termCompGrp>
            <termCompGrp>
              <termComp>der</termComp>
              <termNote type="partOfSpeech">other</termNote>
            </termCompGrp>
            <termCompGrp>
              <termComp>Mitgliedstaat</termComp>
              <termNote type="partOfSpeech">noun</termNote>
              <termNote type="grammaticalNumber">plural</termNote>
            </termCompGrp>
          </termCompList>
          <admin type="status">approved</admin>
          <transacGrp>
                <transac type="transactionType">origination</transac>
                <transacNote type="responsibility">PC</transacNote>
                <date>2014-05-08</date>
          </transacGrp>
        </termGrp>
      </ntig>
    </langSet>
  </body>
</text>
```

Figure 1: An example TBX document.

for a single language. Thus, in order to represent lexical entries in different languages, one lexicon per language needs to be created. In our example, as there are terms for English and German, two lexica need to be created. These lexica contain one lexical entry each, corresponding to the terms *'Zuständigkeit der Mitgliedstaaten'* and *'competence of the Member States'*. The English entry generated from lines 8–14 would look as follows:

```
<http://tbx2rdf.lider-project.eu/data/iate/en>  a  ontolex:Lexicon ;
  ontolex:entry     :competence+of+the+Member+States-en ;
  ontolex:language  "en" .

:competence+of+the+Member+States-en
  a                      ontolex:LexicalEntry ;
  tbx:reliabilityCode    "3"^^xsd:string ;
  tbx:termType           tbx:fullForm ;
  ontolex:canonicalForm  :competence+of+the+Member+States-en#CanonicalForm ;
  ontolex:language       "en" ;
  ontolex:sense          :competence+of+the+Member+States-en#Sense .

:competence+of+the+Member+States-en#CanonicalForm
  ontolex:writtenRep  "competence of the member states"@en .

:competence+of+the+Member+States-en#Sense
  ontolex:reference  :IATE_84.
```

Note that the entry specifies the reliability code (i.e. 3), the type of term (i.e. *full form*), the canonical form (i.e. *'competence of the member states'*), and the language (i.e. *en*). Each lexical entry is assumed to have a `LexicalSense` that represents the meaning of the entry. In this case the meaning is established by `reference` to the terminological concept :IATE_84.

We would generate a similar entry for German, which is identified by the URI `:Zust%C3%A4ndigkeit+der+Mitgliedstaaten-de` and is an entry in the corresponding German lexicon. Note that both entries have a `reference` to `:IATE_84` and are thus cross-lingual equivalents.

So far, we have not yet discussed how composite terms are supposed to be represented. The individual words that make up a term are represented as `constituents` of the composite term. A component is linked to its corresponding lexical entry by way of the `correspondsTo` relation. In the example below, the lexical entry `Zust%C3%A4ndigkeit+der+Mitgliedstaaten-de` is linked to an object `Zust%C3%A4ndigkeit+der+Mitgliedstaaten-de#ComponentList` representing its decomposition via the property `correspondsTo`. This object `Zust%C3%A4ndigkeit+der+Mitgliedstaaten-de#ComponentList` is linked to its components via the property `constituent`. For each component, its part-of-speech and grammatical number (if applicable) are indicated. The decomposition of the German entry for *Zuständigkeit der Mitgliedstaaten* (lines 21–36 in the sample TBX document) is represented in RDF as indicated below:

```
<http://tbx2rdf.lider-project.eu/data/iate/de>  a       ontolex:Lexicon ;
  ontolex:entry     :Zust%C3%A4ndigkeit+der+Mitgliedstaaten-de ;
  ontolex:language  "de" .

:Zust%C3%A4ndigkeit+der+Mitgliedstaaten-de
  a                       ontolex:LexicalEntry ;
  tbx:reliabilityCode     "3"^^tbx:reliabilityCode ;
  tbx:termType            tbx:fullForm ;
  ontolex:canonicalForm  :Zust%C3%A4ndigkeit+der+Mitgliedstaaten-de#CanonicalForm ;
  ontolex:language        "en" ;
  ontolex:sense           :Zust%C3%A4ndigkeit+der+Mitgliedstaaten-de#Sense.

:Zust%C3%A4ndigkeit+der+Mitgliedstaaten-de#CanonicalForm
        ontolex:writtenRep  "Zuständigkeit der Mitgliedstaaten"@de .

:Zust%C3%A4ndigkeit+der+Mitgliedstaaten-de#ComponentList decomp:identifies
    :Zust%C3%A4ndigkeit+der+Mitgliedstaaten-de ;
  decomp:constituent :component1, :component2, :component3 .


:component1 decomp:correspondsTo :Zust%C3%A4ndigkeit-de .
:component2 decomp:correspondsTo :der-de .
:component3 decomp:correspondsTo :Mitgliedstaaten-de .

:Zust%C3%A4ndigkeit-de
  a                       ontolex:LexicalEntry ;
  rdfs:label              "Zuständigkeit"@de ;
  tbx:grammaticalNumber  tbx:singular ;
  tbx:partOfSpeech        tbx:noun.

:der-de
  a                  ontolex:LexicalEntry ;
  rdfs:label         "der"@en ;
  tbx:partOfSpeech  tbx:other.

:Mitgliedstaaten-de
  a                  ontolex:LexicalEntry ;
  rdfs:label         "Mitgliedstaat"@en ;
  tbx:partOfSpeech  tbx:singular ;
  tbx:grammaticalNumber tbx:plural
```

Finally, we discuss how to represent provenance information, in particular that as expressed via transaction elements in TBX. We rely on the PROV ontology[7] for this, as this is the W3C recommended vocabulary to *'represent and interchange provenance information generated in different systems and under different contexts.'* Some provenance information is given on lines 37–42 of Figure 1 and from this we generate the following representation:

```
:Zust%C3%A4ndigkeit+der+Mitgliedstaaten-de
  tbx:reliabilityCode     "3"^^tbx:reliabilityCode ;
  tbx:transaction         :Transaction .

:Transaction
  a                        prov:Activity , tbx:Transaction ;
  tbx:transactionType      "origination"@en ;
  prov:endedAtTime         "2014-05-08"^^<http://www.w3.org/2001/XMLSchema#date> ;
  prov:wasAssociatedWith  :Agent .

:Agent
  a           prov:Agent ;
  rdfs:label  "PC" .
```

[7] http://www.w3.org/TR/prov-o/

3. Application to IATE and EMN

In this section, we describe how IATE and the European Migration Network (EMN) datasets were converted into RDF. Table 1 provides information about the size of the generated RDF resources.

Resource	Size (terms)	RDF Triples
IATE	8,081,142	74,023,248
EMN	8,855	106,283

Table 1: Size of the resources described in this paper (without links)

3.1 Converting IATE to RDF

IATE is the current EU's inter-institutional terminology database and successor of several preexisting databases like EURODICAUTOM (Commission), TIS (Council) and EUTERPE (Parliament), among others. IATE is managed by a management group with representatives from different institutions including the European Parliament, the European Commission, the Council of the European Union, the European Court of Justice, the European Central Bank and the Translation Centre for the Bodies of the European Union. Published in 2007, IATE contains more than 8 million terms in all official 24 EU languages and it is still growing at a pace of 300 new terms added every day[8]. It covers a broad spectrum of domains: politics, law, economics, science, energy, etc. The IATE database can be queried online[9], and the web receives about 3600 visits per hour, with 70 million queries a year.

IATE data exports are available as a single dump file for download on the IATE website[10], or on the EU Open Data Portal[11] and, since February 2015, via the tool IATEExtract that permits choosing the languages of interest[12]. This dump is provided in TBX format, described in the previous section. The TBX data fields used by IATE are very well documented[13] and are fully compatible with the TBX specification. Data is structured in three levels: (i) abstract "concepts" which are language independent, (ii) language level with specific info for each language and (iii) term level. IATE has been integrated in different CAT tools and

[8] According to https://tke2014.coreon.com/slides/2014_06_19_104_1150_Maslias_et_al.pdf

[9] http://iate.europa.eu/

[10] http://iate.europa.eu/tbxPageDownload.do

[11] https://open-data.europa.eu/en/data/dataset/iate

[12] Dealing with a huge files supposes a hurdle for average computer users and translators had found simpler but lengthier manners e.g. http://multifarious.filkin.com/2014/07/13/what-a-whopper/.

[13] http://iate.europa.eu/tbx/IATE%20Data%20Fields%20Explaind.htm

databases[14] (Babelnet, Linguee, MateCat, MemoQ, SDL Trados Studio, DVX2/3, CafeTran), and is also accessible from a Firefox plugin[15], Wordpress widget[16] etc.

We converted the data dump for IATE into RDF using the TBX2RDF converter described below in section 5. Each terminological concept in IATE was transformed into a `skos:Concept`. One lexicon was generated for each of the 24 languages and each term was represented as one lexical entry in the corresponding lexicon. Decomposition and provenance information was represented as described above in Section 2.

3.2 Converting EMN to RDF

The EMN glossary describes terminology for use in the immigration and asylum domain. We implemented a crawler to download the HTML pages for the EMN and implemented an ad-hoc converter directly into *lemon*-based RDF format. It was converted into *lemon* in a manner that follows that of IATE, in that a `Lexicon` was created for each language and then for each of the available terms a `LexicalEntry` was created. The forms of the EMN datasets were preprocessed by removing elements in brackets as well as elements separated from the main term by special characters. In this way we created in total of 338 concepts with 8,855 terms in 22 European languages. Furthermore, we also included a concept definition, semantic relations, explanatory comments and references to other terms.

4. Linking Experiments

In order to link the different terminologies to each other in addition to Babelnet[17], we established links between `skos:Concepts` across datasets by matching the canonical form (lemma) of the corresponding lexical entries in different languages. The number of languages for which the lexical entries for a given concept match, is regarded as an indicator of the quality of the match; that is, the more languages yield a match, the higher the quality of the induced link is expected.

In particular, EMN concepts were linked to IATE concepts by searching for string matches between corresponding EMN lexical entries and IATE lexical entries in multiple languages. In order to improve recall, we used Snowball stemming[18] for the 11 supported EU languages and transformed all strings to lowercase. The search was limited to IATE concepts associated with migration (subject field 2811).

[14] `http://termcoord.eu/iate/download-iate-tbx/iate-data-in-cat-tools-and-databases/` or `http://santrans.net/`

[15] `http://www.maslias.eu/2013/07/iate-european-terminology-database.html?view=classic`

[16] `http://termcoord.eu/resources/`

[17] `http://babelnet.org/`

[18] `http://snowball.tartarus.org/`

Multiple IATE concepts can match a single EMN concept. In order to decide between candidate matches, we counted the number of languages for which each match holds and used this count as a measure for match plausibility. We induced 3,028 links between EMN and IATE by considering all possible matches. Only considering the best match for each EMN concept resulted in 2,038 links (compare Table 2).

Resources	Number of links	Percentage of EMN	Precision
EMN-BabelNet	1,347	15%	69%
EMN-IATE (all matches)	3,082	35%	93%
EMN-IATE (best matches)	2,038	23%	94%

Table 2: Number of links between resources and precision of mapping.

EMN concepts were linked to BabelNet by using Babelfy (Moro et al., 2014), a named entity linking service. Invoking the Babelfy disambiguation algorithm on the written representation of the lexical entries, we extracted all the synsets with which Babelfy annotated the written representation with and considered only those annotations consisting of exactly one synset. A precision of 69% was determined by manually comparing concept definitions for a sample of 100 matches.

On the basis of the existing linking between MASC and BabelNet and the above mentioned induced links between EMN and IATE (3,028, see Table 2) as well as between EMN and BabelNet (1,347, see Table 2), by transitive closure we were able to induce 700 links between IATE and BabelNet (via EMN as pivot), 37,405 links between EMN and MASC (via BabelNet as pivot) and 7,794 between IATE and MASC (via BabelNet and EMN as pivots). The results are summarized in Table 3. To give an example, the EMN term *'visa'* was linked to the matching term associated with IATE concept 3556819 and to BabelNet synset bn:00080087n, which in turn had been used to annotate 15 different tokens in MASC.

Resources	Number of links
IATE-EMN-BabelNet	700
EMN-BabelNet-MASC	37,405
IATE-EMN-BabelNet-MASC	7,794

Table 3: Number of transitive links added to resources.

We evaluated the linking precision by manually evaluating a sample of 100 generated links. Precision of the linking is defined as the number of correctly created links divided by the number of generated links. Precision was determined by manually comparing terms, definitions and sources for a sample of matches. A link was judged as correct if the concepts share the same source or if their definitions do not contradict and there was no better matching concept. The precision of the linking is shown in Table 2. The precision of linking EMN to IATE is quite high, which is due to the fact that they are terminologies and typically only contain one sense or meaning for a certain term / lexical entry. In contrast, BabelNet contains many possible senses for each lexical entry, so that the meaning needs to be actually disambiguated automatically, which is an error-prone process. We evaluated the precision of the induced links in dependence of the number of languages for which the written representations matched. This analysis is shown in Figure 2 and Table 4. We observe that there is a clear improvement when considering links induced when the written representations for more than five languages match.

Languages	Matches	Precision
1–5	669	82%
6–10	448	95%
11–15	846	97%
16–20	992	96%

Table 4: Number of EMN-IATE mappings by number of languages matching.

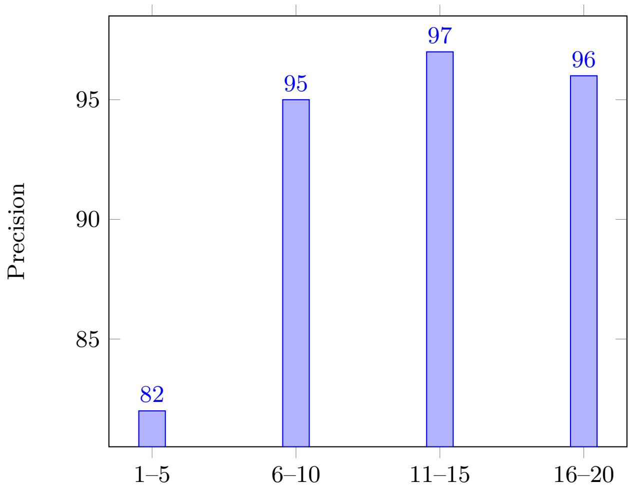

Figure 2: Precision of linking by number of languages matching for EMN-IATE mapping.

5. TBX2RDF Public Service

With the purpose of disseminating the publication of terminologies as linked data, a TBX2RDF Public Service has been released capable of converting terminologies in TBX to RDF[19]. The online converter consists of a form which accepts a TBX document to be uploaded or directly

[19] http://tbx2rdf.lider-project.eu/

pasted in a box, and produces the RDF counterpart. Additional mappings can be added for specific flavours of TBX. The converter can be invoked in strict mode, in which case strict adherence to the TBX standard is ensured[20], and lenient mode, where some tolerance is applied. Additional information is shown when the TBX document does not conform to the standard, or when unexpected input is found. This demonstrative application has been key for gathering feedback on the quality of the conversion and the usefulness of the project itself.

In addition, the TBX2RDF Public Service is offered as a HTTP REST service[21], supporting its integration with existing applications. The service can be tested online[22] and it is accessible through its endpoint, offering the three following main functionalities:

- **Translate:** This is the basic conversion service, which admits as parameters the input TBX document, the desired namespace assigned to the new RDF resources, the option that forces the parser to have strict behaviour (optional) and an alternative set of mappings (optional). The service returns either the RDF document or an error message with a description of the problems encountered, if any.
- **ReverseTranslate:** This functionality is not yet fully implemented in the service. The goal is to admit the input RDF document as input together with a set of optional mappings and return the corresponding TBX document.
- **Enrich:** This functionality is not yet fully implemented in the service. The goal is to admit as input the URL of a terminology published as linked data and to return links to other terminologies as result.

6. Conclusion

In this paper, we have presented a new approach to publishing and linking terminologies using Linked Data principles. We have briefly described the advantages of applying linked data principles to terminologies and presented a model for representing terminologies in RDF. This model has been applied to the transformation of two terminologies, IATE and EMN, into Linked Data. We have also presented an approach to link terminologies to each other automatically. A public service for converting terminologies in TBX format to RDF has been implemented as part of this work and is freely available for anyone wanting to convert their terminologies into linked data. Future work involves developing better algorithms for linking as well as extending the current converter from TBX to RDF by a roundtrip functionality as well as by a service that can enrich existing terminologies with links to other terminologies.

In addition, following the creation (i.e., conversion) and harmonisation (i.e., linking) of open terminologies like IATE and EMN, we advance our work in a practical application of

[20] Conformance of the XML document to the DTD can be validated through the TBX Checker `http://www.tbxconvert.gevterm.net/`

[21] `http://tbx2rdf.lider-project.eu/converter/doc`

[22] `http://tbx2rdf.lider-project.eu/converter/tbx2rdf.html`

RDF-represented terminologies in industry/business-related scenarios. We have been experimenting with Tilde Terminology[23]) already. Finally, in collaboration with the H2020-funded FREME innovation action[24], the next step is the application of linked data terminologies within real world business cases. The FREME project builds an open innovative commercial-grade framework of e-services for semantic and multilingual enrichment of digital content. The FREME project is developing enrichment services by building on existing mature semantic and multilingual technologies and cloud-based infrastructures previously developed by partners and used in business value adding components. The integration of the TBX2RDF service as a further component is currently planned.

7. Acknowledgements

This work is supported by the European projects LIDER (FP7 610782) and FREME (Horizon2020 644771) as well as by the Spanish Ministry of Economy and Competitiveness (project TIN2013-46238-C4-2-R and a Juan de la Cierva grant) and the German Research Foundation (Cluster of Excellence Cognitive Interaction Technology 'CITEC', EXC 277, at Bielefeld University).

8. References

Cabrio, E., Aprosio, A. P. & Villata, S. (2014). These are your rights: A natural language processing approach to automated RDF licenses generation. In *The Semantic Web: Trends and Challenges*, Springer, pp. 255–269.

Heath, T. and Bizer, C. (2011). *Linked Data: Evolving the Web into a Global Data Space.* Synthesis Lectures on the Semantic Web. Morgan & Claypool Publishers.

ISO (2008). Systems to manage terminology, knowledge and content – TermBase eXchange (TBX).

McCrae, J., Spohr, D. & Cimiano, P. (2011). Linking lexical resources and ontologies on the semantic web with lemon. In *The Semantic Web: Research and Applications - 8th Extended Semantic Web Conference, ESWC 2011, Heraklion, Crete, Greece, May 29-June 2, 2011, Proceedings, Part I*, pp. 245–259.

Moro, A., Cecconi, F. & Navigli, R. (2014). Multilingual word sense disambiguation and entity linking for everybody. In *Proceedings of the 13th International Conference on Semantic Web.*

Rodriguez-Doncel, V., Villata, S. & Gomez-Perez, A. (2014). A dataset of RDF licenses. In Hoekstra, R., editor, *Proceedings of the 27th International Conference on Legal Knowledge and Information System*, pp. 187–189.

[23] `http://www.tilde.com/term`

[24] `http://www.freme-project.eu`