Marseille, 11–16 May 2020

From Linguistic Resources to Ontology-Aware Terminologies: Minding the Representation Gap

Giulia Speranza, Maria Pia di Buono, Johanna Monti and Federico Sangati
UNIOR NLP Research Group
University of Naples L'Orientale
{gsperanza, mpdibuono, jmonti, fsangati}@unior.it

Abstract

Terminological resources have proven crucial in many applications ranging from Computer-Aided Translation tools to authoring softwares and multilingual and cross-lingual information retrieval systems. Nonetheless, with the exception of a few felicitous examples, such as the IATE (Interactive Terminology for Europe) Termbank, many terminological resources are not available in standard formats, such as Term Base eXchange (TBX), thus preventing their sharing and reuse. Yet, these terminologies could be improved associating the correspondent ontology-based information. The research described in the present contribution demonstrates the process and the methodologies adopted in the automatic conversion into TBX of such type of resources, together with their semantic enrichment based on the formalization of ontological information into terminologies. We present a proof-of-concept using the Italian Linguistic Resource for the Archaeological domain (developed according to Thesauri and Guidelines of the Italian Central Institute for the Catalogue and Documentation). Further, we introduce the conversion tool developed to support the process of creating ontology-aware terminologies for improving interoperability and sharing of existing language technologies and data sets.

Keywords: Terminology, Linguistic Resources Conversion, Semantic Enrichment

1. Introduction

The availability of terminological resources is crucial for experts in different fields; for instance, translators can integrate standard terminology in Term Base eXchange (TBX) format into their Computer-Aided Translation (CAT) tools, terminologists and linguists can reuse termbases in authoring tool, Natural Language Processing (NLP) experts may use them for several applications, mainly multilingual and cross-lingual ones. One of the major problems for experts who do not possess specific knowledge of terms and phraseology of a domain is to detect and use specialised terms normally used by domain-experts. In addition, the choice of the corresponding translations for terms is not always straightforward because of polysemy and ambiguity problems.

Terminologies, developed by linguists, terminologists and domain experts, represent essential resources for language-based applications and platforms especially those connected with CAT and authoring tools. Terminology creation and maintenance are tasks that determine the quality of the final product of a translation process. Good-quality and controlled terminology is critical for the success of a translation process, whether it is a human or an automatic one.

Terminological databases have, indeed, an important role in translation technology, such as Machine Translation (MT), and in many multilingual applications as well, such as Multilingual Information Retrieval (MLIR), Cross-language Information Retrieval (CLIR) applications among others. The IT sector should consider the importance of terminology in the translation process and integrate it into the tools.

At the same time, terminological resources should be made available in standard formats so that they can be used extensively in different applications, from CAT tools and MT to NLP.

An interesting example in this respect is provided by the IATE (Interactive Terminology for Europe) Termbank[1], a concept-oriented database covering more than 100 subject fields, which has been used in the EU institutions and agencies since summer 2004 for the collection, dissemination and management of EU-specific terminology[2]. IATE is a main reference not only for the EU institutions but also for Language Service Providers (LSPs) and freelance translators and it is for this reason that the IATE resources have been made available in TBX for free download and integration in translation tools.

Sometimes terminological resources, developed by domain experts, are not available in a standard format and therefore cannot be used in many applications. In addition, several specialized glossaries and thesauri are created and maintained by experts and professional figures working in the respective domains of knowledge, who might not be aware of the linguistic potential of those resources. Consequently, many domain professionals do not take into consideration the advantages of storing such terminologies according to standards and enriching them with ontological information. There is therefore the need to standardize these resources so that they can be aligned with similar resources in other languages and, subsequently, be used in translation technology and multilingual applications.

Taking these issues into account, this contribution describes the process and the methodologies adopted in the automatic conversion into TBX of linguistic resources together with their semantic enrichment based on the formalization of ontological information into terminologies.

We present a proof-of-concept of our methodology using the Italian Linguistic Resource for the Archaeologi-

[1] https://iate.europa.eu/home

[2] IATE represents a crucial tool in the translation process of the European Union since it guarantees the consistent use of terms and the possibility to have access to a multilingual repository with reliable information. The database is constantly updated and it currently counts 7,996,776 terms.

cal domain (developed according to Thesauri and Guidelines of the Italian Central Institute for the Catalogue and Documentation (ICCD)). Further, we introduce the conversion tool developed to support the process of creating ontology-aware terminologies for improving interoperability and sharing of existing language technologies and data sets.

Section 2 describes related work in the field of terminology management and the tools developed to convert from different formats to TBX. Section 3 presents Term Base eXchange (TBX) the international standard (ISO 30042:2019) for the representation of structured concept-oriented terminological data. Our approach in developing ontology-aware terminologies in the domain of archaeology is explained in Section 4, including a description of the proposed semantic mapping with the TBX format (section 4.1), the presentation of the linguistic resources used as proof-of-concept of our approach (section 4.2), and finally the conversion process (section 4.3). In Section 5 we describe how our approach can support interoperability in the development of common Language Technologies and how it can be integrated into the existing infrastructures. Sections 6 concludes the paper.

2. Related Works

As several scholars pointed out (Wright et al., 2010; Melby, 2012), terminology management is often a heterogeneous activity involving different formats, data models and practices with people inside and outside the industry showing a strong tendency to store terminology in simple formats such as CSV and spreadsheets.

Previous works committed to the adoption of TBX as a standard format for the creation and exchange of terminology, focused on proposing alternatives to commercial terminology management systems in order to create terminology standardization.

Melby (2008) provided an open source conversion tool for porting MRC (Multiple Rows per Concept) Term Table format into TBX-Basic. MRC files can be created in either a text editor or a spreadsheet thus not requiring knowledge or expertise in XML-based languages.

Wright et al. (2010) proposed a TBX dialect (Glossary-TBX) which is specifically envisioned for representing basic glossaries and it is accompanied by a conversion tool which enables conversion between UTX-Simple, GlossML, the TBX family, and OLIF. On the basis of this work, the informal TBX steering committee created another dialect, namely TBX-Min, for representing very minimal and simple termbases, such as spreadsheets.

In order to promote the use of TBX-Min, Lommel et al. (2014) proposed a tool for converting existing glossaries stored in spreadsheet formats into TBX-Min. A selection of tools[3] is as well provided by the informal TBX steering committee.

Stanković et al. (2014) developed a wizard integrated in their terminological information system for converting termbases into TBX. Pinnis et al. (2013) developed a cloud-computing platform called TaaS (Terminology as a Service) to provide management of existing terminological data which also supports TBX export functionalities.

As previously stated, TBX is also the standard format chosen for the downloadable version of IATE repository. IATE, a central terminology database for all the institutions, agencies and other bodies of the European Union, provides a single access point to the existing European terminological resources, besides an infrastructure for the constitution, shared management and dissemination of these resources (Johnson and Macphail, 2000). With a current total number of 935K entries, 7.1 MM terms and 26 languages[4], this database represents the reference in the terminology field, and is considered to be the largest multilingual terminology database in the world.

Finally, previous line of research on terminology insisted on establishing a connection between terminology and formal ontology within the same domain, thus shading the light on the double nature of terminology (conceptual and linguistic) and stressing how useful such combination is, both for translators and experts in the field as well as Natural Language Processing (NLP) applications (Roche, 2012; Moreno and Pérez, 2000; Navigli and Velardi, 2008).

3. Terminological Data Representation

TermBase eXchange (TBX) is an international standard (ISO 30042:2019)[5] for the representation of structured concept-oriented terminological data. Initially published by the Localization Industry Standards Association (LISA), it has been released under a Creative Commons license in 2011, when LISA ceased its operations.

The foundations for TBX have been established by three international standards: (i) TMF (ISO 16642:2003), which defines the structural metamodel for TBX and other TMLs (terminological markup languages); (ii) ISO 12620, which provides an inventory of data-categories for terminological data; (iii) MARTIF (ISO 12200:1999), which presents the basis for the core structure of TBX and the XML styles of its elements and attributes.

TBX provides an XML-based framework to manage terminology, knowledge and content, by means of several processes, such as analysis, descriptive representation, dissemination, and interchange (exchange).

The TBX framework is composed of two main modules: a core-structure module and an XCS (eXtensible Constraint Specification) module. The former includes high-level elements which are in correspondence with the TMF metamodel. The latter is based on a formalism for identifying a set of data-categories and their constraints. The core-structure module is defined in a DTD used together with an XCS file that applies additional data-category constraints.

Data-categories are the result of the specification of a given data field, e.g., part of speech, or grammatical number. In order to guarantee high interoperability, TBX provides a default set of data-categories that are commonly used in terminological databases. Data-categories can be imple-

[3] http://tbxconvert.gevterm.net/tbx-min/

[4] https://iate.europa.eu/download-iate

[5] For this documentation we refer to the official documentation available at https://www.gala-global.org/sites/default/files/uploads/pdfs/tbx_oscar_0.pdf

mented using either an attribute or the content of an element.
A data-category implemented using an attribute is a terminological data-category that is defined according to ISO 12620, such as */definition/*, and one that is specified as a value of the name attribute in the default XCS file.
A data-category implemented as the content of an element is a simple data-category, that is, one value of a closed set of values (pick-list). These terminological data-categories are also documented according to ISO 12620.
The specification of the value of an attribute, the content of an element, or one or more structural levels, may be formalized through data-category constraints, which limit the application of a meta data-category, a core-structure module data-category that takes a type attribute and facilitates modularity. The default TBX data-categories and their constraints includes elements or attribute, implemented directly in the core-structure DTD, and specializations, e.g., concept relations, properties and description of terms, of the metadata-categories.

4. Ontology-Aware Terminologies

Our approach has a twofold goal: supporting the development of terminological resources in standard formats and integrating information from domain-specific ontologies into such resources. The use of an ontology in the upgrading of these resources may ensure knowledge sharing, maintenance of semantic constraints, semantic ambiguities solving, and inferencing on the basis of ontology concept networks. The integration of ontological prescriptions will enhance terminologies, adding information and constraints useful for providing elements from logical semantics, which can be described as truth-conditional semantics and model-theoretic semantics.
Currently, most existing representation models used to formalize semantics in RDF, namely vocabularies developed on the basis of OntoLex-Lemon (McCrae et al., 2017) (e.g., PreMon (Rospocher et al., 2019), Framester (Gangemi et al., 2016), REO (Brown et al., 2017)), address lexical semantic aspects, which capture the underlying predicate-argument structure. The emerging need is formalizing propositions, as idealised sentence suitable for logical manipulation, so that the meaning of the various parts of the propositions are given by a group of interpretation functions which license important inferences. In order to achieve this goal, terminological resources should contain domain-specific ontology information, suitable for providing a description which combines lexical and logical aspects. Combining lexical and logical aspects in the development of terminological resources could improve semantic inference which relies on logical representations, furthering generalizations and supporting formal semantics for logical operators within linguistic theories.
To this aim, we propose a conceptual mapping that is source agnostic and language independent, which means that can be applied to all linguistic resources and languages.

4.1. Semantic Mapping

Linguistic and semantic information stored in LRs need an adequate TBX representation, capable of preserving the information themselves, guaranteeing a consistent semantic representativeness and a high interoperability. TBX format provides all the elements needed for such a conceptual mapping, as shown in Table 1. In fact, it is structured hierarchically onto 3 levels, namely concept level, language level and term level, which can be used for a complete description of ontology-aware terminologies.

Concept Level. At the concept level, which is language-independent, it is possible to specify the domain of knowledge covered in the linguistic resource, using the TBX `subjectField` data category. There is also the possibility of going deeper in the domain and indicate the different subdomain levels, hierarchically dependent on the general `subjectField`, specifying a `metaType` associated to the `subjectField` with `<subjectField metaType="Level1">`. Thus, the information stored into `subjectField` and its `metaType` can be used to describe a fine-grained taxonomy of the knowledge domain and its subdomains, useful in order to restrict the application focus.
Furthermore, TBX allows to make an explicit cross-reference to a resource (URI, URL, or local file path) external to the TBX file at the concept level; therefore, it is possible to specify the reference to an ontology stored as external resource by means of External Cross-Reference `<xref>` pointing to the ontology URI/URL itself, e.g., `http://www.cidoc-crm.org/cidoc-crm`. This link to such an external resource is highly important since it establishes a connection between the terminological resource and an ontological reference by means of a persistent URI. The ontological reference guarantees semantic interoperability between different sources, thus enriching information with external semantic data specifically related to the field.
Finally, it is possible to supply a term definition using `<descrip type="definition">`. Such an information helps non experts in the technical field in representing and framing the meaning and use of a specific entry in relation to a particular subject or activity.

Language Level. At the language level, the specific language of the entries can be indicated, with compliance to the language code taken from ISO 639-1, ISO 639-2, or ISO 639-3. Including the language indication in the `<langSet>` field represents a good practice in the development of termbases.

Term Level. Each entry in the LR corresponds to the TBX data category `<term>`, which is a language specific representation of a concept in a given domain or subject field, thus pertaining to the term level. At the term level it is possible to specify whether a term is a single word, grouping it with `<tig>`, or a multiword expression (MWE), using the `<ntig>` nesting. In order to decompose the MWE into its single components, the `<termComp>` element can be used. For both single words and MWEs, it is possible a further specification of their part of speech, which can be represented in TBX with `<termNote type="partOfSpeech">`, adding a value indicated in the pick-list (i.e., noun, verb, adjective, adverb, properNoun, other). The POS indication is particu-

larly useful in order to disambiguate possible homographs.
Term morpho-syntactic information, such as gender and number, can be specified in TBX by means of <termNote type="grammaticalGender"> and <termNote type="grammaticalNumber">. Making the grammatical information explicit is useful also for agreement in construction at the syntactic level.
The aforementioned linguistic information are enriched by semantic information using ad-hoc types of the TBX termNote element. In fact, by means of <TermNote type = "hypernyms"> it is possible to indicate broader categories the term belongs to. Viceversa, <TermNote type = "hyponyms"> allows to include narrower and more specific categories.
The possibility of including these information allows to introduce the representation of the existing IS-A relationships among terms. By means of <termNote type="hypernyms"> and <termNote type="hyponyms"> it is possible to include in the LR the representation of semantic relations about terms.
A resource containing such information as contained in WordNet[6] sysnsets' model (Miller, 1995; Fellbaum, 2010), together with the representation of their variants and synonyms, may prove (i) effective in the alignment process with other already existing resources, and (ii) useful when gathering external information from other resources.
Further types of termNote are created to specify term variants, as intended by the ISO 12620[7], namely alternative forms of a term such as spelling variants or different capitalization. To this aim, the Term Type value "variant" has been introduced. The use of such information allows to improve the assessment of terminological harmonisation and consistency at an intra-textual level.
Finally, in order to express synonyms, which are terms that represent the same or a very similar concept as the main entry (ISO 12620), one can use the <termNote type="synonym">.
To keep the ontology reference also at the term level, we decided to use an external cross-reference by means of <xref> pointing to an URI which refers to a specific ontological class used to represent that term, e.g., for the term *dynos con anse ad anello* the <xref> value will be http://www.cidoc-crm.org/cidoc-crm/E22_Man-Made_Object[8].
Additional information may be specified, such as a reliability code by means of <descrip type="reliabilityCode">, authorship and authors' roles by means of <transacGrp> and an example of sentence containing the word in context by means of <descrip type="context">[9].

[6] https://wordnet.princeton.edu/

[7] ISO 12620:1999(E)

[8] This URI refers to the standard ontology used in the Cultural Heritage domain, i.e., CIDOC Conceptual Reference Model (CRM). See Section 4.2 for more information.

[9] Such elements, are not presented in Table 1, as they are not involved into the conceptual mapping process.

4.2. Case Study: ICCD thesauri and its LRs

In order to provide a proof-of-concept of this approach to simplify the process of creating ontology-aware terminologies, transforming linguistic resources into TBX files, we use a sample of data from the Italian Linguistic Resources for the Archaeological domain (di Buono et al., 2014), built according to the Lexicon-Grammar descriptive method, form part of the DELA System[10].
These LRs, firstly developed to be used into NooJ enviroment (Silberztein, 2015), are tool independent and include information taken from the Thesauri and Guidelines of the Italian Central Institute for the Catalogue and Documentation (ICCD)[11]. ICCD resources are organized in several thesauri and dictionaries, which present different levels of information granularity. The most informative one is the Object definition dictionary which provides, for each entry, information about the Broader Term [BT], Broader Term Partitive [BTP1], Broader Term Partitive [BTP2], Narrower Term [NT], Narrower Term Partitive [NTP], Use [USE], Use For [UF].
BT and NT fields indicate a taxonomic classification, so that broader and narrower terms can be related to the main entry. For instance, *amuleto* (amulet) is an element of both *Strumenti, Utensili e Oggetti d'uso* (Tools), which is a general category, and *Amuleti e oggetti per uso cerimoniale, magico e votivo* (Magic & Votive Supplies), which is a specific category.
The NTP, BTP1 and BTP2 fields specify the lemma in its partitive uses at both narrower and broader levels. This specification helps to infer that *amuleto* occurs in different MWE entries, for instance: *amuleto a forma di anatra* (duck amulet), *amuleto a forma di ariete* (ram amulet) and so on.
UF is a non-preferential lemma (i.e., a variant); this implies that *cornetto* (horn amulet) can stand for *amuleto* (and its specific types), but ICCD guidelines suggest to use the first one. All possible variants are lemmatized, including those having even a low-frequency use.
The electronic dictionary, used in our proof-of-concept, is composed of 11000 entries, with both simple words and MWEs, including spelling variants, i.e., (*dinos+dynos+déinos*) *con anse ad anello* (ringed-handle (dinos+dynos+déinos)), and synonyms, generally extracted from the UF field, i.e., *kylix a labbro risparmiato* (spared-lip kylix), which stands for lip cup or *cratere* (crater) which stands for *vaso* (vase).
These LRs have been improved adding ontological information to incorporate more information than thesauri. Indeed, with reference to a thesaurus, an ontology also stores language-independent information and semantic relations. The ontological reference for the Archaeological domain resources are provided by the ICOM International Committee for Documentation (CIDOC) Conceptual Reference Model (CRM)[12] (Doerr, 2003), an ISO standard since

[10] Dictionnaires Électroniques du LADL (Laboratoire d'Automatique Documentaire et Linguistique)

[11] http://www.iccd.beniculturali.it/index.php?it/240/vocabolari

[12] http://www.cidoc-crm.org/

Levels	Input LRs	Output TBX
Concept Level	General domain and sublevels	<descrip type="subjectField">...</> <subjectField metaType="Level1">...</>
	General Ontology Reference	<xref type="externalCrossReference"> <target="external_id">...</>
	Definition	<descrip type="definition">...</>
Language Level	Language	<langSet xml:lang="xx">...</>
Term Level	Single Word - Entry	<tig>...</> <term>...</>
	MWE - Entry	<ntig>...</> <term>...</> <termComp>...</>
	Ontology Class Reference	<xref type="externalCrossReference"> <target="external_id">...</>
	Synsets	<termNote type="hypernyms">...</> <termNote type="hyponyms">...</>
	Category	<termNote type="partOfSpeech">...</>
	MWE - Internal Structure	<termNote type="partOfSpeech">...</>
	Inflectional Information	<termNote type="grammaticalNumber"> <termNote type="grammaticalGender">
	Variants	<termNote type="variant">...</>
	Synonyms	<termNote type="synonym">...</>

Table 1: Conceptual mapping between LRs information and TBX data categories

2006, compatible with the Resource Description Framework (RDF). CIDOC CRM provides definitions and a formal structure for describing the implicit and explicit concepts and relationships used in Cultural Heritage documentation.

Thus, for each entry, e.g., *dinos con anse ad anello*, there is a specification on its POS (Category), internal structure to represent POS of each component in a multiword expression, and inflectional code (FLX), referring to inflectional information[13]; its variants (VAR), and synonyms (SYN), if any. Furthermore, with reference to the ICCD prescriptions, the pertaining knowledge domain together with its taxonomic classification (DOM)[14] and a reference to the specific CIDOC CRM Class (CCL) are specified.

4.3. Conversion Process

To support the development of ontology-aware terminologies, and, more in general, the development of terminologies, both based on the conversion of existing LRs into TBX files, we create a converter suitable for automatically mapping these resources with standard elements[15]. In order to simplify the conversion process for non-expert users we also propose an application, available in the form of a chatbot on the Telegram platform under the name CSV2TBX[16], to lead users in the use of our tool. The conversion process through the CSV2TBX chatbot is achieved by a pre-set group of conversational instructions and the user only has to follow them while interacting with the chatbot in the chatting platform.

Before uploading the CSV file containing the LR to be converted, the chatbot asks users to provide general information about the resources which their intend to use. In other words, users have to define the concept and language levels, namely the language, e.g., Italian, the subjectField, (which refers to the domain of knowledge the terminology belongs to, e.g., Archaeology[17]), the ID prefix, (a unique identifier associated to each entry, which is a combination of characters and numbers e.g., AR_001[18]) and whether or not they want to include the name and the URL/URI of a reference domain ontology, (e.g., CIDOC CRM and `http://www.cidoc-crm.org/cidoc-crm` respectively). These indications specified during the interaction performed in the chatting platform are stored and mapped in the TBX file as well.

The CSV structure to upload may contain 9 fields, of whom just the first two are mandatory, separated by 8 semicolumns (;). These fields indicate: i) the term, either a single or multiword expression; ii) the POS, e.g., N; iii) the internal POS for MWEs, in the form of a sequence of POS for each component, e.g., NPNPN); iv) the grammatical info, e.g., gender and number in the form of ms-+; v) the variants, i.e., orthographical variations of the term; vi) the synonyms, to represent terms conveying a very similar concept as their respective term entry; vii) a definition, namely a brief explication of the term as a dictionary gloss; viii) the hypernyms, as hierarchically higher and

[13]Such codes refers to inflectional grammars developed in the DELA system.

[14]This taxonomic relation is directly derived from the ICCD taxonomy and represented in our LRs through an alphanumeric value.

[15]The converter, released under a CC license, is available at `https://github.com/unior-nlp-research-group/TBX-Converter`

[16]`https://t.me/CSV2TBX_bot`

[17]It is worth stressing that IATE, the EU's terminology database, and EuroVoc, a multidisciplinary thesaurus covering the activities of the EU, apply a numeric identifier for subject domains and their sub-levels.

[18]The user only has to indicate the first two characters since the progressive number is automatically set by the converter.

more general lexical entries comprising the term; ix) the ontology class which represents the entry in the form of its alpha-numeric identifier in the corresponding domain ontology e.g., E22_Man-Made_Object.
Before uploading the file, an automatic message will specify which CSV fields are mandatory and which ones are optional, so that the users can easily check their LR file. In cases of wrong input during the guided procedure, the chatbot will send a warning message and support users to provide the right input.
Once the input has been received by the chatbot, the file is sent to the converter tool. Subsequently, the converter automatically maps the LR data fields to the TBX data categories and gives as output a TBX file, enriched with ontological information. Figure 1 shows the TBX output for the entry *dinos con anse ad anello*, formalized into a CSV file as follows:

```
dinos con anse ad anello;N;NPNPN;ms-+;
dynos con anse ad anello,déinos
con anse ad anello;;...;RA1SUOCR;
E22_Man-Made_Object[19]
```

During the conversion process, all the linguistic and semantic information are preserved.
The `termEntry` specifications hold the information about the subject field and a description for the entry. As the entry is a MWE, each component of the entry is further specified as single element composing the full form, with its own lexical and morpho-syntactic information.
Furthermore, if the LR file stores ontological information as well, the converter provides an automatic URL/URI creation by means of a simple users' specification about the ontology which the resources refer to. In fact, users are asked whether or not they want to specify the referring ontology and the information will be inserted into the `xref` field as URI/URL. Since the ontological information are represented at both concept level and term level, two types of `xref` can be specified. The first one refers to the general ontology for a specific domain, i.e., CIDOC CRM for the Archaeological domain, and the second one stands for the specific ontology class for each term, i.e., E22_Man-made_Object, in our example.
Finally, a linguistic variant, which refers to the same concept, and two hypernyms, which stand for the IS-A relationships of our entry, are represented. The output file, being developed according to terminological standards, may be easily integrated into CAT and MT and authoring tools and submitted to an evaluation process.
The converter has been developed as an application which simplifies the integration into third-party tools. As we will discuss in the following section, the capability of being integrated into existing language infrastructures supports the development of common language technologies.

[19]Due to lack of space we removed from this example the definition information which is formalized in the TBX output in Figure 1.

5. Interoperability and Language Technologies

The European Parliament resolution of 11 September 2018 on language equality in the digital age[20] provides some recommendations on creating a European language technology (LT) platform for sharing of services and enabling and empowering European SMEs to use LTs. To address these recommendations, several initiatives and projects aim at providing tools supporting interoperability and sharing of existing language technologies and data sets, e.g. the European Language Grid (ELG)[21], Prêt-à-LLOD[22], Elexis[23].
In order to contribute to the development of common language technologies and support these sharing initiatives, we plan to share our TBX resources and integrate our service into existing language infrastructures, i.e., ELG.
ELG intends to establish the primary platform for LTs in Europe, involving several stakeholders from the language technology sector to create a community which shares technologies and data sets through its platforms and deploys them through the grid and connect with other resources. ELG deals with several content, namely services, language resources, data sets, tools, directory content. Our converter could contribute to both functional content and non-functional content, as the ELG platform provides an easy and efficient way for LT providers to create and upload containers and linguistic resources (Rehm, 2019). Our converter can be uploaded and integrated into other systems, as it can be realised by containerising the LT service (dockerisation) by means of a set of API descriptors, suitable to advertise all capabilities of this service.

6. Conclusion

In this work, we presented our approach to transform linguistic resources into ontology-aware terminologies. Such an approach relies on a precise conceptual mapping of linguistic and semantic information to TBX elements. Besides the linguistic information stored in these resources, we rely on an enrichment 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 this way, our research intends to contribute to semantic-aware language services which exploit the already existing LRs and enhance the development of new one as well. Nowadays, several specialized glossaries and thesauri are created and maintained by experts and professional figures working in the respective domains of knowledge, who might not be aware of the potential of those resources. Consequently many domain professionals do not take into consideration the advantages of storing such terminologies according to standards and enriching them with ontological information. On the other side, it has been reported that, even inside the terminology industry, there still are some practitioners collecting terminolo-

[20]http://www.europarl.europa.eu/doceo/document/TA-8-2018-0332_EN.html
[21]https://www.european-language-grid.eu/
[22]https://www.pret-a-llod.eu/
[23]https://elex.is/

```
...
    <termEntry id="RA_286">
      <descripGrp>
        <descrip type="subjectField">Archaeology</descrip>
        <descrip type="definition">Recipiente a larga bocca rotonda, sagomata da un
         basso orlo verticale derivato da prototipi bronzei; essendo privo di piede,
         aveva bisogno di essere collocato su un tripode o su un sostegno sagomato.
         Serviva per mescere vino e acqua.</descrip>
        <xref type="URI"target= "http://www.cidoc-crm.org/cidoc-crm">CIDOC CRM Ontology</xref>
      </descripGrp>
    </termEntry>
    <langSet xml:lang="it">
      <ntig>
        <termGrp>
          <term>dinos con anse ad anello</term>
          <termNote type="partOfSpeech">noun</termNote>
          <termNote type="grammaticalGender">masculine</termNote>
          <termNote type="grammaticalNumber">singular</termNote>
          <xref type="URI" target="http://www.cidoc-crm.org/cidoc-crm/E22_Man-Made_Object">
           CIDOC CRM Class</xref>
          <termCompList type="lemma">
            <termCompGrp>
              <termComp>dinos</termComp>
              ...
            </termCompGrp>
            <termCompGrp>
              <termComp>con</termComp>
              ...
            </termCompGrp>
            <termCompGrp>
              <termComp>anse</termComp>
              ...
            </termCompGrp>
            <termCompGrp>
              <termComp>ad</termComp>
              ...
            </termCompGrp>
            <termCompGrp>
              <termComp>anello</termComp>
              ...
            </termCompGrp>
          </termCompList>
          <termNote type="variant">déinos con anse ad anello</termNote>
          <termNote type="hypernyms01">Contenitori e recipienti</termNote>
          <termNote type="hypernyms02">Strumenti - utensili - oggetti d'uso</termNote>
        </termGrp>
      </ntig>
    </langSet>
...
```

Figure 1: Example of TBX output for the entry *dinos con anse ad anello*

gies in simpler file formats and people struggling with the complexities posed by terminological standards (Wright et al., 2010; Lommel et al., 2014).

Having easy access to a converter tool for the automatic transformation and ontological enrichment of these LRs supports: (i) the creation of trustworthy resources made by domain experts or reliable and authoritative organizations and bodies; (ii) interoperability and sharing of existing resources; (iii) the development of an interconnected cloud of ontology-aware terminologies.

As future work, we intend to enhance our converter towards supporting its containerization and integration with other ontology-aware services, e.g., Terme-à-LLOD (di Buono et al., Forthcoming).

7. Acknowledgements

This work has been partially supported by Programma Operativo Nazionale Ricerca e Innovazione 2014-2020 - Fondo Sociale Europeo, Azione I.2 “Attrazione e Mobilità Internazionale dei Ricercatori” Avviso D.D. n 407 del 27/02/2018 and by POR Campania FSE 2014-2020 “Dottorati di Ricerca a Caratterizzazione Industriale”.

Authorship contribution is as follows: Giulia Speranza is author of Section 2 and 4.1; Maria Pia di Buono is author of Section 3, 4 and 4.2; Johanna Monti is author of Section 1 and 6. Federico Sangati is author of Section 4.3 and 5.

8. Bibliographical References

Brown, S., Bonial, C., Obrst, L., and Palmer, M. (2017). The rich event ontology. In *Proceedings of the Events and Stories in the News Workshop*, pages 87–97.

di Buono, M. P., Monteleone, M., and Elia, A. (2014). Terminology and knowledge representation. italian linguistic resources for the archaeological domain. In *Proceedings of Workshop on Lexical and Grammatical Resources for Language Processing*, pages 24–29.

di Buono, M. P., Cimiano, P., Elahi, M. F., and Grimm, F. (Forthcoming). Terme-à-llod: Simplifying the conversion and hosting of terminological resources as linked data. In *Submitted to LREC 2020*.

Doerr, M. (2003). The cidoc conceptual reference module: an ontological approach to semantic interoperability of metadata. *AI magazine*, 24(3):75–75.

Fellbaum, C. (2010). Wordnet. In *Theory and applications of ontology: computer applications*, pages 231–243. Springer.

Gangemi, A., Alam, M., Asprino, L., Presutti, V., and Recupero, D. R. (2016). Framester: a wide coverage linguistic linked data hub. In *European Knowledge Acquisition Workshop*, pages 239–254. Springer.

Johnson, I. and Macphail, A. (2000). Iate-inter-agency terminology exchange: development of a single central terminology database for the institutions and agencies of the european union. In *Workshop on Terminology resources and computation*.

Lommel, A., Melby, A., Glenn, N., Hayes, J., and Snow, T. (2014). Tbx-min: a simplified tbx-based approach to representing bilingual glossaries.

McCrae, J. P., Bosque-Gil, J., Gracia, J., Buitelaar, P., and Cimiano, P. (2017). The ontolex-lemon model: development and applications. In *Proceedings of eLex 2017 conference*, pages 19–21.

Melby, A. K. (2008). Tbx-basic translation-oriented terminology made simple. *Tradumàtica: traducció i tecnologies de la informació i la comunicació*, (6).

Melby, A. K. (2012). Terminology in the age of multilingual corpora. *The Journal of Specialised Translation*, 18:7–29.

Miller, G. A. (1995). Wordnet: a lexical database for english. *Communications of the ACM*, 38(11):39–41.

Moreno, A. and Pérez, C. (2000). Reusing the mikrokosmos ontology for concept-based multilingual terminology databases. In *LREC*.

Navigli, R. and Velardi, P. (2008). From glossaries to ontologies: Extracting semantic structure from textual definitions.

Pinnis, M., Gornostay, T., Skadiņš, R., and Vasiļjevs, A. (2013). Online platform for extracting, managing, and utilising multilingual terminology. In *Proceedings of the Third Biennial Conference on Electronic Lexicography, eLex 2013*, pages 122–131.

Rehm, G. (2019). European language grid: An overview. In *META FORUM, Brussels, Belgium* `https://www.european-language-grid.eu/wp-content/uploads/2019/10/00-03-ELG-Overview-Georg-Rehm.pdf`.

Roche, C. (2012). Ontoterminology: how to unify terminology and ontology into a single paradigm. lrec 2012. In *Eighth International Conference on Language Resources and Evaluation. Istanbul*, pages 21–27.

Rospocher, M., Corcoglioniti, F., and Aprosio, A. P. (2019). Premon: Lodifing linguistic predicate models. *Language Resources and Evaluation*, 53(3):499–524.

Silberztein, M. (2015). *La formalisation des langues: l'approche de NooJ*. ISTE Group.

Stanković, R., Obradović, I., and Utvić, M. (2014). Developing termbases for expert terminology under the tbx standard. *Editors Gordana Pavlović Lažetić Duško Vitas Cvetana Krstev*.

Wright, S. E., Rasmussen, N., Melby, A. K., and Warburton, L. (2010). Tbx glossary: a crosswalk between termbase and lexbase formats. In *Proceedings of developing, updating and coordinating technologies, dictionaries and lexicons for terminological consistency workshop*.