Terme-à-LLOD: Simplifying the Conversion and Hosting of Terminological Resources as Linked Data

Maria Pia Di Buono[1], **Philipp Cimiano**[2], **Mohammad Fazleh Elahi**[2], **Frank Grimm**[2]
[1]UNIOR NLP Research Group, University of Naples L'Orientale
[2]Semantic Computing Group, Bielefeld University
mpdibuono@unior.it
{cimiano, melahi, grimm}@cit-ec.uni-bielefeld.de

Abstract

In recent years, there has been increasing interest in publishing lexicographic and terminological resources as linked data. The benefit of using linked data technologies to publish terminologies is that terminologies can be linked to each other, thus creating a cloud of linked terminologies that cross domains, languages and that support advanced applications that do not work with single terminologies but can exploit multiple terminologies seamlessly. We present Terme-à-LLOD (TAL), a new paradigm for transforming and publishing terminologies as linked data which relies on a virtualization approach. The approach rests on a preconfigured virtual image of a server that can be downloaded and installed. We describe our approach to simplifying the transformation and hosting of terminological resources in the remainder of this paper. We provide a proof-of-concept for this paradigm showing how to apply it to the conversion of the well-known IATE terminology as well as to various smaller terminologies. Further, we discuss how the implementation of our paradigm can be integrated into existing NLP service infrastructures that rely on virtualization technology. While we apply this paradigm to the transformation and hosting of terminologies as linked data, the paradigm can be applied to any other resource format as well.

Keywords: Linguistic Linked Open Data, Terminological Resources, NLP services

1. Introduction

Terminological resources, mainly termbases, represent a core source of data for translation and localization (Stanković et al., 2014). Further, they have important applications in text mining as they provide concepts with which elements of text can be tagged for semantic normalization as well as semantic indexing (Witschel, 2005). Therefore, many intermediate representations of terminological information and tools for termbase management have been developed so far with the main goal of improving the portability and interoperability of those resources. Among the representations of terminological information, the TermBase eXchange (TBX) format has become a standard for terminology information exchange. Such an exchange plays an important role in ensuring consistency and contributes to terminology production and quality through interactive validation processes (joh,).

In recent years, there has been interest in publishing terminological resources as linked data in order to improve interoperability and reuse and a number of approaches proposing to use linked data principles to publish terminologies have been proposed (Cimiano et al., 2015; Rodriguez-Doncel et al., 2015; Montiel-Ponsoda et al., 2015).

The general benefits of publishing language resources as linked data have been described by Chiarcos et al. (2013). In short, the benefit of using linked data technologies to publish terminologies is that terminologies can be linked to each other, in order to create a cloud of linked terminologies that cross domains, languages and that support advanced applications that do not work with single terminologies but can exploit multiple terminologies seamlessly (see the work of Montiel et al. (2015) on integrating two terminologies, TERMACT and Terminesp, using Linked Data). Along these lines, a number of projects have published specific guidelines on how to publish terminological resources using linked data and Semantic Web technologies. For example, as a result of the EC-funded LIDER project[1] and as part of the work of the W3C community group on Best Practices for Multilingual Linked Open Data (BPMLOD)[2], guidelines have been released describing how to publish terminologies in TBX format as linked data using the ontolex-lemon model (McCrae et al., 2011; McCrae et al., 2015). More recently, the Linked Heritage Project[3] has released recommendations for how to manage terminologies in the framework of the Semantic Web.

Yet, a fundamental problem remains, that is that implementing all these guidelines and recommendations is challenging as one needs a detailed understanding of the corresponding vocabularies in addition to technical understanding of data models (e.g., RDF) as well as how to host linked data at a server level. We present a new approach that we call Terme-à-LLOD (TAL), aiming to fill this gap and simplifying the task of converting a terminological resource in TBX format into a linked data resource and ease the task of hosting the linked data resource in such a way that i) URIs resolve, ii) the resource can be browsed, and iii) a SPARQL endpoint is offered. This new paradigm for transforming and publishing standardized terminological resources as linked data relies on a virtualization approach. The approach rests on a pre-configured virtual image of a server that can be downloaded and installed. In our approach we rely on Docker[4], but any other virtualization environment can be used instead.

The remainder of the paper is organized as follows: Section

[1] http://www.lider-project.eu/lider-project.eu/index.html
[2] https://www.w3.org/2015/09/bpmlod-reports/multilingual-terminologies/
[3] http://linkedheritage.eu/
[4] https://www.docker.com

2. describes our new approach and its virtualization process to simplify the transformation and hosting of terminological resources. Following this, Section 3. presents two case studies comprising the transformation of the Inter-Agency Terminology Exchange (IATE)[5] terminology as well as a sample of termbases provided by the Centrum Voor Terminologie (CvT)[6] at Ghent University into Linked Data. Finally, Section 4. shows how our approach can be integrated into existing NLP service infrastructures that rely on virtualization technology such as European Language Grid (ELG)[7] and Teanga[8]. Section 5. discusses related work and how our approach can be integrated into other NLP service frameworks. The paper ends with the conclusion and future work.

2. Terme-à-LLOD Approach

Terme-à-LLOD is a new virtualization paradigm for easing the process of transforming terminological resources into RDF and hosting them as linked data. The virtualization paradigm relies on three main components (Figure 2): a converter (A), a Virtuoso Server[9] (B) (Erling and Mikhailov, 2010; Erling, 2012), and a container (C).

The converter element managing the automatic format transformation is based on the TBX2RDF service[10] (Cimiano et al., 2015) developed by the LIDER project. TBX2RDF maps TBX inputs, including TBX public dialects, i.e., TBX-Core, TBX-Min and TBX-Basic, into RDF format, reusing a set of classes and properties from existing linked open data vocabularies (e.g., OntoLex-Lemon[11]). An example of converting TBX to RDF is can be seen in Figure 2.

The converter produces an RDF output which serves as input to a Virtuoso server, the second component of the TAL virtualization technology. Once the RDF output has been uploaded, the pre-installed server, which hosts the service, exposes the converted data through an endpoint which allows to access them. The server also provides a SPARQL endpoint to other services.

The third element of the virtualization technology is a Docker container that can be easily installed on any Docker environment. The Docker container allows to bundle components, libraries and configuration files of the TAL service and to run the service on different computing environments. Once the container is installed and instantiated, the terminological resource can be pushed via HTTP/Advanced Message Queuing Protocol (AMQP) request to the TBX2RDF converter. Subsequently, the TAL service invokes the transformation to Linked Data using the converter and hosts the resulting RDF as linked data together with a SPARQL endpoint.

The benefit of such a virtualization approach is that the owner of a terminology can easily publish the terminology as linked data without the need to understand the underlying vocabularies in detail nor of the RDF data model or about how to set up a linked data server. Yet, the data remains under full control and can be published under a namespace to represent ownership and provenance.

```
<termEntry id="IATE-84">
  <descripGrp>
    <descrip type="subjectField">1011</descrip>
  </descripGrp>
  <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="da">
    <tig>
      <term>medlemsstatskompetence</term>
      <termNote type="termType">fullForm</termNote>
      <descrip type="reliabilityCode">3</descrip>
    </tig>
  </langSet>
  <langSet xml:lang="nl">
    .....
  </langSet>
  <langSet xml:lang="es">
    .....
  </langSet>
  .....
</termEntry>
```

```
<http://webtentaclel.techfak.uni-bielefeld.de/tbx2rdf_iate/
competence+of+the+Member+States-en>
      a ontolex:LexicalEntry ;
      dct:language <http://www.lexvo.org/id/iso639-3/eng> ;
      tbx:reliabilityCode 3 ;
      <http://www.lexinfo.net/ontology/2.0/lexinfo#termType>
            <http://www.lexinfo.net/ontology/2.0/lexinfo#fullForm> ;
      <http://www.w3.org/ns/lemon/lime#language>
            "en" ;
      ontolex:canonicalForm <http://webtentaclel.techfak.uni-bielefeld.de/
      tbx2rdf_iate/data/iate/competence+of+the+Member+States-en#CanonicalForm> ;
      ontolex:sense <http://webtentaclel.techfak.uni-bielefeld.de/
      tbx2rdf_iate/data/iate/competence+of+the+Member+States-en#Sense> .

<http://webtentaclel.techfak.uni-bielefeld.de/tbx2rdf_iate/
data/iate/dal%C4%ABbvalstu+kompetence-lv#Sense>
      a ontolex:LexicalSense ;
      ontolex:isLexicalizedSenseOf :IATE-84 .

<http://webtentaclel.techfak.uni-bielefeld.de/tbx2rdf_iate/
data/iate/competence+of+the+Member+States-en#CanonicalForm>
      ontolex:writtenRep "competence of the Member States"@en .

<http://webtentaclel.techfak.uni-bielefeld.de/tbx2rdf_iate/
data/iate/medlemsstatskompetence-da#CanonicalForm>
      ontolex:writtenRep "medlemsstatskompetence"@da .
.......
```

Figure 1: TBX (top) to RDF (bottom) conversion

2.1. Virtualization Process

The virtualization technology is contained into a preconfigured virtual image that can be hosted in a corresponding environment consisting of virtual machines communicating with each other over standard protocols. All capabilities of a TAL service are advertised in an OpenAPI descriptor file. This lets consumers discover how to communicate with the service and what result values to expect.

The TAL service automatically gathers the latest version of the TBX2RDF service from GitHub and installs it in a multi-stage container build that makes knowledge of the underlying Java development stack transparent to the end user. TAL adds a Node.js application behind a nginx reverse proxy for HTTP communication with the service. This application is used to orchestrate the different internals of the container and monitor the status or health of the container.

The service is initially provided with term glossaries as standardized TBX files, as defined by ISO standard

[5] https://iate.europa.eu
[6] http://www.cvt.ugent.be
[7] https://www.european-language-grid.eu/
[8] https://teanga.techfak.uni-bielefeld.de/
[9] https://virtuoso.openlinksw.com/
[10] http://tbx2rdf.lider-project.eu/converter/
[11] https://www.w3.org/2016/05/ontolex/

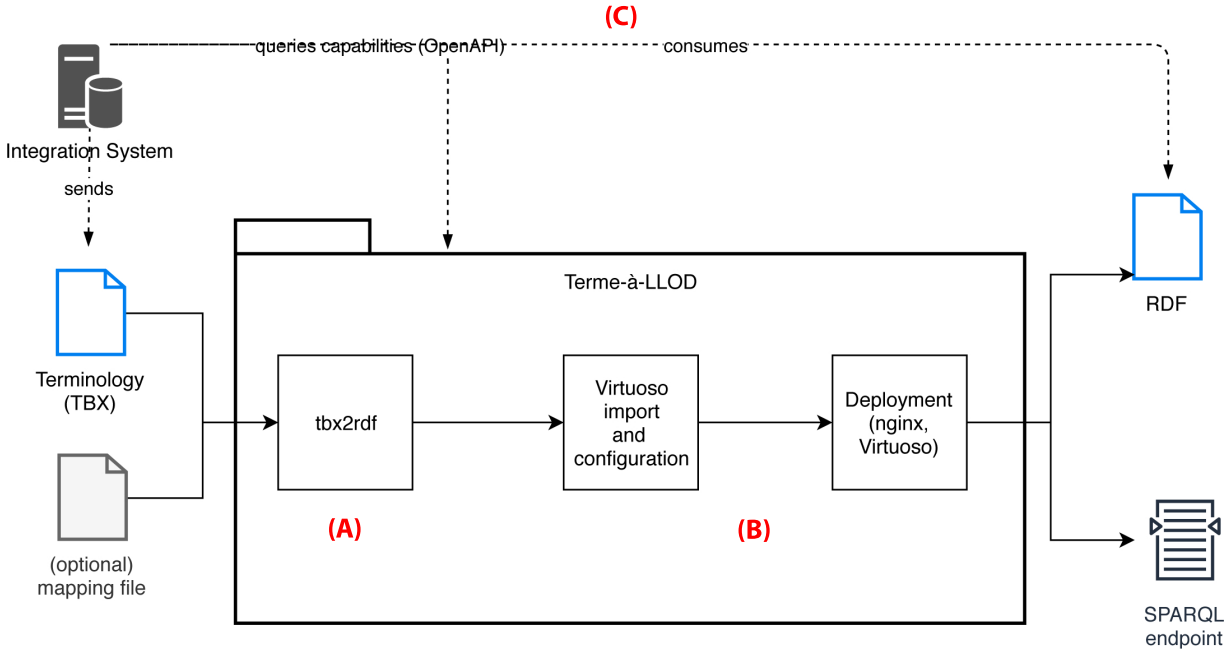

Figure 2: TAL virtualization paradigm

30042:2008, as well as an optional file containing mapping information between TBX and the desired target RDF namespace. A stack of open source software is then used to convert and deploy the glossary data in queryable linked open data formats, namely as a resource description framework (RDF) download and a SPARQL endpoint to query this data.

The container bundles the TBX2RDF converter, implemented as a Java program that reads in the document and builds the DOM tree. The DOM tree is traversed and elements are mapped to appropriate object-oriented datastructures. These datastructures are then serialized as RDF.

The resulting RDF is exposed to a web server for further downstream tasks that require the full dataset and automatically populates an instance of the Virtuoso open source database engine[12].

Since terminology databases can be of considerable size, the container additionally exposes a status application that allows an end user to monitor the conversion progress and status of each service instance. During regular use, the ecosystem issues a new instance of the TAL Docker container that is available on GitHub[13]. It is either initialized as an empty instance or provided with the state or database content of a previously created instance.

The only knowledge required to setup a service instance is minimal and generally regards the specific ecosystem used to work with the service. Specialized knowledge of TBX, LLOD or NLP technologies is not necessary and abstracted away in order to make these resources more approachable.

In fact, a Docker container is a lightweight, standalone, executable package ("container image") of software that can be seen as a template to bootstrap everything required to run an application: code, runtime, a lean operating system, system libraries and settings. The Docker engine (e.g., pure Docker, Kubernetes or a platform-as-a-service cloud offering) enables containerized applications to run anywhere consistently on essentially any infrastructure. A Docker volume is used to retain the results of costly conversion processes across updates and reboots. During bootup the TAL container starts a Node.js application and nginx web server. The service is immediately discoverable through the OpenAPI descriptor. The conversion process itself has to only be run once, e.g., by the party maintaining a particular terminology. Subsequent users can consume from the initialized service instance by either post-processing the generated RDF artifact that is exposed via HTTP or querying the SPARQL endpoint that hosts the resulting linked data structures.

3. Use Cases: IATE and GENTERM

In order to provide a proof-of-concept of this approach to simplify the process of transforming terminological resources into RDF and hosting the RDF as linked data, we used a sample of data from two sources. The data source is the Inter-Agency Terminology Exchange (IATE) repository and the second are a number of termbases hosted by the Centrum Voor Terminologie (CvT) at Ghent University.

3.1. IATE

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 (joh,). With a current total number of 935K entries, 7.1 MM terms and 26 languages[14], this database represents the reference in the terminology field, and is considered to be

[12] http://vos.openlinksw.com/owiki/wiki/VOS

[13] https://github.com/ag-sc/terme-a-llod and https://hub.docker.com/r/agsc/terme-a-llod

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

TBX field	RDF element
TBX resource	`void:Dataset`
Term	`skos:Concept`
Langset	`ontolex:Lexicon`
TIG/NTIG	`ontolex:LexicalEntry`
TermGrp	`ontolex:canonicalForm`
TermCompList	`ontolex:decomp`
TermCompGrp	`decomp:correspondsTo`
DescrGrp	properties of the lexical entry
TransGrp/Transaction	`tbx:Transaction`

Table 1: Conceptual mapping of TBX fields and RDF elements.

TBX input	Runtime	# Terms	# Triples	# Lang
IATE	25.2m	5851035	52603182	25
Pharmaceutical*	5.2s	4629	71347	2
Diseases*	3.0s	799	12650	2
Waste management*	2.5s	396	6109	2
Solar energy*	2.9s	205	3758	2
Printmaking*	2.5s	223	3426	2

Table 2: Information about IATE and GENTERM conversion process (Entries marked with * are courtesy of GENTERM).

the largest multilingual terminology database in the world. Data, provided in TBX format and made available without a copyright protection, can be freely downloaded and reproduced, for personal use or for further non-commercial or commercial dissemination[15].

3.2. GENTERM

The second sample of data has been extracted from the termbases developed by the Centrum Voor Terminologie (CvT) - GENTERM[16]. The center, active within the Department of Translation, Interpreting and Communication of Ghent University, co-ordinates the Department's activities on terminology and terminography and makes available a small set of termbases, which are the result of several students' projects. GENTERM termbases belong to different domains (e.g., pharmaceutica, waste management, solar energy, diseases, printmaking). We provide a proof-of-concept of the workins conversion with all these six term bases.

3.3. Transformation to linked data

We converted the IATE and CvT TBX terminologies using our Terme-à-LLOD service and expose test instances on a central demonstration server[17] that can be used in combination with other workflows.

As already mentioned, the conversion process is mainly based on the use of the TBX2RDF converter. Several vocabularies, mainly W3C recommendations, have been used during the conversion process, namely OntoLex-lemon, SKOS, RDF-schema, DCAT, VOID, PROV-O, LIDER TBX Ontology. The TBX fields we consider during the conversion process and the mapping elements selected from aforementioned vocabularies are shown in Table 1. The TBX Resource field is not explicitly represented, as the whole dataset represents the TBX resource. A TBX resource is thus represented as a void:Dataset to which provenance and licensing information can be attached. Furthermore, a langset is not represented as such in the data. Instead, one ontolex:Lexicon is created for each language for which a LangSet is defined. The collection of all the terms for a given language will belong to the corresponding language-specific ontolex:Lexicon. The DescrGrp field contains descriptions of the term or context that are mapped to appropriate properties of the lexical entry or the context. A general overview of the conversion process is available in Table 2. For each termbasis used in the conversion process, we present the runtime needed, the number of terms stored in the termbasis, the number of triples resulting in the output files, and the number of languages we converted. Figure 3 shows an example of TAL final output, namely the exposure of an RDF terminological resource which can be browsed to access more specific information about each term.

3.4. Linking IATE and GENTERM

After the conversion process, we established links among the different termbases tested in the use case by means of Simple Knowledge Organization System (SKOS)[18] concepts. The linking across GENTERM and IATE terminologies has been accomplished matching the corresponding lexical entries in different languages by means of a

[15] `https://iate.europa.eu/legal-notice`

[16] `http://www.cvt.ugent.be/downloads.htm`

[17] `http://scdemo.techfak.uni-bielefeld.de/termeallod/`

[18] SKOS is a vocabulary for representing knowledge organization systems (KOS), such as thesauri, classification schemes, subject heading and taxonomies in RDF.

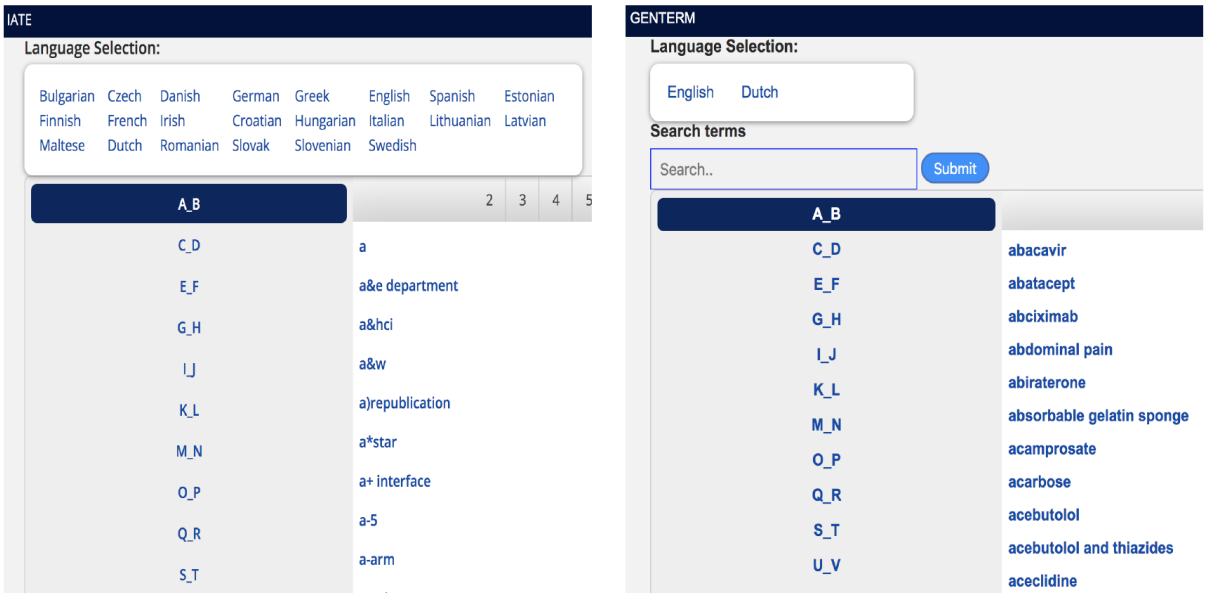

Figure 3: Example of converted terminological resources exposed in TAL service.

string comparison based on `skos:exactMatch`. This match allows linking, for instance, the term *nefopam*, from the Pharmaceutical termbasis in GENTERM, to the corresponding term in IATE, which has an alpha-numeric identifier, i.e., IATE-3545983. Once links among the terminologies have been established, users can explore a term across all the converted and exposed termbases.

Table 3 shows the number of links between GENTERM and IATE. Even though the GENTERM terminology covers different domains in two languages (English and Dutch), the termbases available are very small in comparison to IATE. This explains the low number of links for some of the proposed domains, e.g., GENTERM Solar energy-IATE in Table 3.

4. Integration into language infrastructures

Language infrastructures represent one of the leading areas for the digital economic growth[19], as well as a key element to enable an inclusive Digital Single Market[20]. Several initiatives and projects aim at providing tools supporting interoperability and sharing of existing language technologies and data sets. In order to contribute to the development of common language technologies and support these sharing initiatives, as a proof-of-concept we have developed an approach to integrate our Terme-à-LLOD approach into two language infrastructures, namely Teanga and ELG. Furthermore, such an integration proves the interoperability of our approach which relies on virtualization services.

Teanga is a linked data based platform for natural language processing (NLP) which enables the use of many NLP services from a single interface (Ziad et al., 2018). Many platforms have been developed to improve the interoperability among different NLP services and, consequently, reduce the manual effort required to process data. One of the main issues in these proposals is the need of following specific standards to develop NLP services which interoperate smoothly on the platform. To address this issue, Teanga uses linked data and open, semantic technologies to describe the boundaries between different services in terms of application programming interface (API) descriptors. Given an API endpoint, these descriptors can be used to automatically discover the capabilities of a particular service and specify data types of possible inputs and outputs to the particular NLP service.

ELG is an initiative to establish the primary platform for language technologies (LT) in Europe (Rehm et al., Forthcoming). Its main goal is involving several stakeholders from the language technology sector to create a community which shares technologies and data sets through the platforms, deploys them through the grid and connects them with other resources. ELG addresses some of the recommendations identified in the European Parliament resolution of 11 September 2018 on language equality in the digital age, namely creating a European LT platform for sharing of services and enabling and empowering European SMEs to use LTs.

The grid platform has been built with robust, scalable, reliable, widely used technologies that are constantly developed further. It presents the ability to scale with the growing demand and supply of resources through an interactive modern web interface, providing the base technology for a catalogue or directory of functional services, data sets, tools, technologies, models and LT companies, research organisations, research projects, service and application types, languages. ELG deals with several types of content, that is, services, language resources, data sets, tools and directory content. The TAL service presented in this paper belongs to the type functional content, comprising of containerized services that can be uploaded and integrated into other systems. To support this integration, the

[19] `https://www.tractica.com/research/natural-language-processing/`

[20] `https://ec.europa.eu/commission/priorities/digital-single-market_en`

Termbases	Lang	Number of Links
GENTERM Pharmaceutical-IATE	English	1380
	Dutch	1084
GENTERM Diseases-IATE	English	22
	Dutch	27
GENTERM Waste management-IATE	English	114
	Dutch	109
GENTERM Solar energy-IATE	English	12
	Dutch	20
GENTERM Printmaking-IATE	English	35
	Dutch	21

Table 3: Results from the linking process.

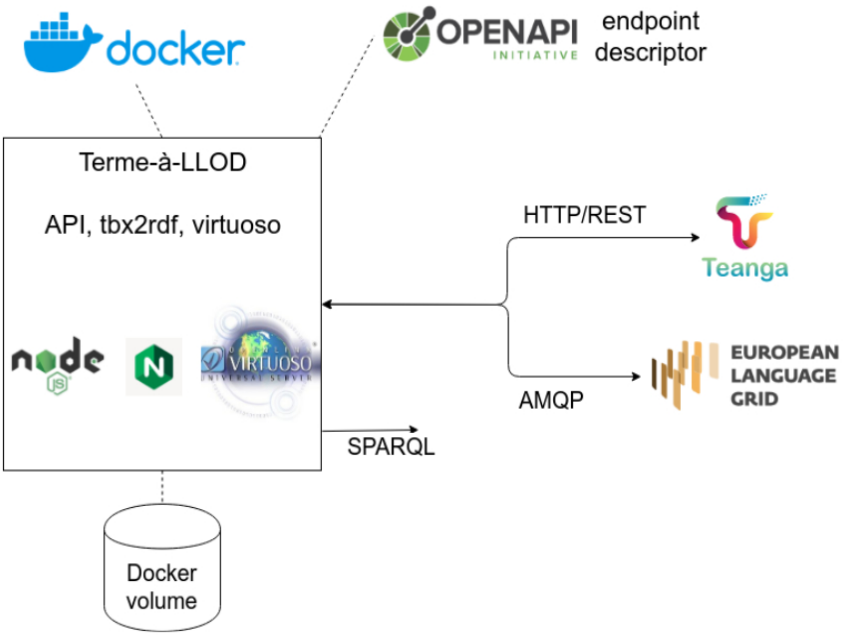

Figure 4: Integration of TAL service into language infrastructures

ELG platform provides an easy and efficient way for LT providers to create and upload containers (Rehm, 2019).
The integration of our TAL service into Teanga and ELG (Figure 4) enables the ecosystem to incorporate data from external terminology glossaries as linked data. Such an integration, based on the development of a docker container, increases the usability of our service, providing a principled way for integrating our service into other applications consuming the terminological data. The ELG cluster and the workflow of Teanga application, namely the integration servers, can query the OpenAPI descriptor in TAL service. When used as a service endpoint, the service automatically connects to a message queue via the AMQP that acts as a service bus for inter-service communication and lets other software consume data.
In recent years, containerization replaced many use cases where the only previous option was full virtualization of a system. Especially in modern, service-based, architectures virtualization often implied many of the same hardships as maintaining a system on bare metal machines, including operating system maintenance and the ability to scale such a system still being costly and subject to the underlying virtualization infrastructure. By introducing containers, most of these aspects of software deployments are abstracted into purely operational questions and become practically transparent to the developer. The containerization of the TAL service ensures portability across systems by abstracting the underlying hardware following a virtualization approach while at the same time supporting efficient deployment of the application.

5. Related Work

The TermBase eXchange (TBX) format has become an international standard (ISO 30042:2019)[21] for exchange of terminological information. It allows the representation of structured concept-oriented terminological data providing

[21] 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`

an XML-based framework to manage terminology, knowledge and content by means of several processes, such as analysis, descriptive representation, dissemination, and interchange (exchange). OntoLex-Lemon has been proposed early on as a Linked Data format for representing terminological resources (Cimiano et al., 2015). Ontolex-lemon has been applied to the conversion of Terminesp into Linked Data (Bosque-Gil et al., 2015) as well as to the transformation of a set of freely available terminology databases from the Catalan Terminological Centre, TERMCAT (Montiel-Ponsoda et al., 2015).

Guidelines for converting TBX data to Linked Data have been developed as part of the LIDER project[22]. These guidelines explain how the TBX data model can be mapped to the Ontolex-lemon model and provides a step-by-step example how TBX data can be transformed into RDF following this mapping.

An alternative set of recommendations and guidelines have been developed by the Linked Heritage and Athena projects (Leroi et al., 2011). The document proposes a three-step methodology to digitalize terminologies for publication in the Semantic Web consisting of three three steps: i) conceive your terminology, ii) make your terminology interoperable, iii) link your terminology to a network. As data-model, the document proposes to use the SKOS model rather than the Ontolex-lemon model.

The approach proposed in this paper has focused on the transformation of terminological resources; yet, the principled approach of simplifying the work of transforming resources into RDF would apply to other data formats as well. There has been some work on transforming lexicographic resources as well as WordNets into Linked Data using lemon-Ontolex (McCrae et al., 2012; Eckle-Kohler et al., 2015; Ehrmann et al., 2014; McCrae et al., 2014). There has been work on transforming corpora into RDF (Chiarcos and Fäth, 2017). The approach described here could be applied to those data formats as well.

We have integrated our transformation component into the Teanga (Ziad et al., 2018) and ELG (Rehm et al., Forthcoming) infrastructrues. There are other NLP architecures into which TAL container could be integtated.

WebLicht[23] is an environment for building, executing, and visualizing the results of NLP pipelines, which is integrated into the CLARIN infrastructure (Hinrichs and Krauwer, 2014). NLP tools are implemented as web services that consume and produce the Text Corpus Format (TCF)[24], an XML format designed for use as an internal data exchange format for WebLicht processing tools. It ensures semantic interoperability among all WebLicht tools and resources by defining a common vocabulary for linguistic concepts in TCF schema. The services and resources are developed as web services in the CLARIN framework. The services are exposed using metadata descriptions Component Metadata Infrastructure (CMDI) [25]. CMDI describes functionalities offered by a service, pre and postconditions, and specifications of data that is consumed and produced by service.

The Language Application (LAPPS)[26] (Ide et al., 2014) Grid is a framework that provides access to NLP processing tools and resources and enables pipelining these tools to create custom NLP applications, as well as access to language resources such as mono- and multilingual corpora and lexicons that support NLP. The semantic interoperability of language services is achieved by the Web Services Exchange Vocabulary (Ide et al., 2016), which specifies terminology for a core of linguistic objects and features exchanged by services. Recently, the services are deployed in the cloud using Docker images. While we have integrated our TAL service into ELG and Teanga as a proof-of-concept, it could also be integrated into the WebLicht enviroment as well as LAPPS Grid following the same principles.

6. Conclusion

We have proposed a virtualization approach to support the conversion and hosting of terminologies as linked data. The approach can in principle be applied to any language and lexical resource beyond terminologies using the same principles.

We have demonstrated the applicability of our approach via the conversion into RDF and hosting as linked data of six terminologies in total: the well-known IATE termbase and five smaller termbases hosted by Ghent University. A public Docker container has been implemented and is free available for everyone wanting to convert and host their terminologies. We have described the integration of our approach into state-of-art language infrastructures, namely Teanga and ELG.

Within the European project Prêt-à-LLOD[27], which focuses on making linguistic data ready to use by the use of state-of-the-art technologies, in particular linked data, a further integration of this service is currently planned. Prêt-à-LLOD aims at creating a new methodology for building data value chains applicable to a wide-range of sectors and applications and based around language resources and language technologies that can be integrated by means of semantic technologies. The Terme-à-LLOD approach proposed here follows the aims of the EC-funded Prêt-à-LLOD project of providing *Linked-data based NLP services as data* so that they are sustainable and can be readily used and deployed by third parties. In future work we will enhance the architecture and implementation of the TAL service towards supporting linking any other linked data compliant terminology to the one hosted by TAL.

7. Acknowledgements

This work has been funded by project Prêt-à-LLOD which is funded under European Union's Horizon 2020 research and innovation programme under grant agreement No. 825182.

[22] https://www.w3.org/2015/09/bpmlod-reports/multilingual-terminologies/

[23] https://weblicht.sfs.uni-tuebingen.de/weblichtwiki/index.php/Main_Page

[24] https://weblicht.sfs.uni-tuebingen.de/weblichtwiki/index.php/The_TCF_Format

[25] https://www.clarin.eu/content/component-metadata

[26] http://www.lappsgrid.org/

[27] https://www.pret-a-llod.eu

8. Bibliographical References

Bosque-Gil, J., Gracia, J., Aguado-de Cea, G., and Montiel-Ponsoda, E. (2015). Applying the ontolex model to a multilingual terminological resource. In Fabien Gandon, et al., editors, *The Semantic Web: ESWC 2015 Satellite Events*, pages 283–294, Cham. Springer International Publishing.

Chiarcos, C. and Fäth, C. (2017). CoNLL-RDF: Linked corpora done in an nlp-friendly way. In *Language, Data, and Knowledge - First International Conference, LDK 2017, Galway, Ireland, June 19-20, 2017, Proceedings*, pages 74–88.

Chiarcos, C., McCrae, J., Cimiano, P., and Fellbaum, C. (2013). Towards open data for linguistics: Linguistic linked data. In *New Trends of Research in Ontologies and Lexical Resources*, pages 7–25. Springer.

Cimiano, P., McCrae, J. P., Rodríguez-Doncel, V., Gornostay, T., Gómez-Pérez, A., Siemoneit, B., and Lagzdins, A. (2015). Linked terminologies: applying linked data principles to terminological resources. In *Proceedings of the eLex 2015 Conference*, pages 504–517.

Eckle-Kohler, J., McCrae, J. P., and Chiarcos, C. (2015). lemonuby - A large, interlinked, syntactically-rich lexical resource for ontologies. *Semantic Web*, 6(4):371–378.

Ehrmann, M., Cecconi, F., Vannella, D., McCrae, J. P., Cimiano, P., and Navigli, R. (2014). Representing multilingual data as linked data: the case of babelnet 2.0. In *Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC)*, pages 401–408.

Erling, O. and Mikhailov, I. (2010). Virtuoso: RDF support in a native RDBMS. In *Semantic Web Information Management*, pages 501–519. Springer.

Erling, O. (2012). Virtuoso, a hybrid rdbms/graph column store. *IEEE Data Eng. Bull.*, 35(1):3–8.

Hinrichs, E. and Krauwer, S. (2014). The CLARIN research infrastructure: Resources and tools for ehumanities scholars. In *Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC 2014)*, Reykjavik, Iceland, may. European Language Resources Association (ELRA).

Ide, N., Pustejovsky, J., Cieri, C., Nyberg, E., Wang, D., Suderman, K., Verhagen, M., and Wright, J. (2014). The language applications grid. In *Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC 2014)*, Reykjavik, Iceland, may. European Language Resources Association (ELRA).

Ide, N., Suderman, K., Verhagen, M., and Pustejovsky, J. (2016). The language applications grid web service exchange vocabulary. In *Revised Selected Papers of the Second International Workshop on Worldwide Language Service Infrastructure WLSI (2015)*, pages 18–32, Kyoto, Japan. Springer-Verlag New York, Inc.

). iate.

Leroi, M.-V., Holland, J., and Cagnot, S. (2011). *Your terminology as a part of the semantic web recommendations for design and management*. Repro Stampa Ind. Grafica, Villa Adriana Tivoli.

McCrae, J., Spohr, D., and Cimiano, P. (2011). Linking lexical resources and ontologies on the semantic web with lemon. In *Extended Semantic Web Conference*, pages 245–259. Springer.

McCrae, J., Montiel-Ponsoda, E., and Cimiano, P. (2012). Integrating wordnet and wiktionary with lemon. In *Linked Data in Linguistics*, pages 25–34. Springer.

McCrae, J., Fellbaum, C., and Cimiano, P. (2014). Publishing and linking wordnet using lemon and RDF. In *Proceedings of the 3rd Workshop on Linked Data in Linguistics*.

McCrae, J. P., Cimiano, P., and Doncel, V. R. (2015). Guidelines for linguistic linked data generation: Multilingual terminologies (tbx). *https://www.w3.org/2015/09/bpmlod-reports/multilingual-terminologies/*.

Montiel-Ponsoda, E., Bosque-Gil, J., Gracia, J., de Cea, G. A., and Vila-Suero, D. (2015). Towards the integration of multilingual terminologies: an example of a linked data prototype. In *Proceedings of the conference Terminology and Artificial Intelligence (TAI)*, pages 205–206.

Rehm, G., Berger, M., Elsholz, E., Hegele, S., Kintzel, F., Marheinecke, K., Piperidis, S., Deligiannis, M., Galanis, D., Gkirtzou, K., Labropoulou, P., Bontcheva, K., Jones, D., Roberts, I., Hajic, J., Hamrlova, J., Kacena, L., Choukri, K., Arranz, V., Mapelli, V., Vasiļjevs, A., Anvari, O., Lagzdiņš, A., Meļņika, J., Backfried, G., Dikici, E., Janosik, M., Prinz, K., Prinz, C., Stampler, S., Thomas-Aniola, D., Gomez Perez, J. M., Silva, A. G., Berrio, C., Germann, U., Renals, S., and Klejch, O. (Forthcoming). European language grid: An overview. In *Submitted to LREC 2020*.

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`.

Rodriguez-Doncel, V., Santos, C., Casanovas, P., Gómez-Pérez, A., and Gracia, J. (2015). A linked data terminology for copyright based on ontolex-lemon. In *AI Approaches to the Complexity of Legal Systems*, pages 410–423. Springer.

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*.

Witschel, H. F. (2005). Terminology extraction and automatic indexing. *Terminology and Content Development*, page 363.

Ziad, H., McCrae, J. P., and Buitelaar, P. (2018). Teanga: A linked data based platform for natural language processing. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*.