Proceedings of the 12th Conference on Language Resources and Evaluation (LREC 2020), pages 5660–5667
Marseille, 11–16 May 2020

Recent Developments for the Linguistic Linked Open Data Infrastructure

Thierry Declerck[1], John McCrae[2], Matthias Hartung[3], Jorge Gracia[4], Christian Chiarcos[5], Elena Montiel[6], Philipp Cimiano[7], Artem Revenko[8], Roser Sauri[9], Deirdre Lee[10], Stefania Racioppa[1], Jamal Nasir[2], Matthias Orlikowski[3], Marta Lanau-Coronas[4], Christian Fäth[5], Mariano Rico[6], Mohammad Fazleh Elahi[7], Maria Khvalchik[8], Meritxell Gonzalez[9], Katharine Cooney[10]

[1]DFKI GmbH, Germany; [2]National University of Ireland Galway, Ireland; [3]Semalytix GmbH, Germany;
[4] University of Zaragoza, Spain; [5]Goethe University Frankfurt, Germany;
[6]Universidad Politécnica de Madrid, Spain; [7]Bielefeld University, Germany; [8]Semantic Web Company, Austria;
[9]Oxford University Press, United Kingdom; [10]Derilinx, Ireland
[1]{declerck, stefania.racioppa@dfki.de}@dfki.de; [2]{john.mccrae,jamal.nasir}@insight-centre.org
[3]{hartung,matthias.orlikowski}@semalytix.com; [4]{jogracia,mlanau}@unizar.es
[5]christian.chiarcos@web.de, faeth@em.uni-frankfurt.de; [6]{emontiel,mariano.rico}@fi.upm.es
[7]cimiano@cit-ec.uni-bielefeld.de, melahi@techfak.uni-bielefeld.de
[8]{artem.revenko,maria.khvalchik}@semantic-web.com; [9]{Roser.Sauri,Meritxell.Gonzalez}@oup.com
[10]{deirdre,katharine}@derilinx.com

Abstract

In this paper we describe the contributions made by the European H2020 project "Prêt-à-LLOD" ('Ready-to-use Multilingual Linked Language Data for Knowledge Services across Sectors') to the further development of the Linguistic Linked Open Data (LLOD) infrastructure. Prêt-à-LLOD aims to develop 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. We describe the methods implemented for increasing the number of language data sets in the LLOD. We also present the approach for ensuring interoperability and for porting LLOD data sets and services to other infrastructures, as well as the contribution of the projects to existing standards.

Keywords: Linguistic Linked Open Data, Standards, Infrastructure

1. Introduction

Language technologies increasingly rely on large amounts of data. Better access and usage of language resources will enable to provide multilingual solutions that support the further development of language technologies for the emerging Digital Single Market in Europe. However, language data is rarely 'ready-to-use' and language technology specialists spend over 80% of their time on cleaning, organizing and collecting language datasets. Reducing this effort promises huge cost savings for all sectors where language technologies are required.

The Prêt-à-LLOD project[1] aims at increasing the uptake of language technologies by exploiting the combination of linked data and language technologies, that is Linguistic Linked Open Data (LLOD), to create ready-to-use multilingual data (Cimiano et al., 2020). Prêt-à-LLOD is achieving this by 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 project develops tools for the discovery, transformation and linking of datasets which can be applied to both data and metadata in order to provide multi-portal access to heterogeneous data repositories.

Another goal of Prêt-à-LLOD is to automatically analyse licenses in order to deduce how data may be lawfully used and sold by language resource providers. Finally, the project provides tools to combine language services and resources into complex pipelines by use of semantic technologies. This leads to sustainable data offers and services that can be integrated and deployed into various platforms, including as-yet-unknown platforms, and can be self-described with linked data semantics. Our approach is being validated in four pilots, where data value chains are built for pharmaceutical applications, technology providers, and government services.

In the following sections we present briefly the Linguistic Linked Open Data cloud and the OntoLex-Lemon representation model for lexical data, two of the main initiatives upon which Prêt-à-LLOD is building and further developing. We then discuss some of the objectives of the project, before presenting methodologies put in place in order to achieve them and the industrial pilots designed to demonstrate the relevance and applicability of these methods. We sketch the relations that Prêt-à-LLOD has to other infrastructures before describing the contributions made by the project to standardization activities. We close with an outlook for the next steps.

2. Linguistic Linked Open Data Cloud

The Linguistic Linked Open Data (LLOD) cloud[2] is an initiative, which was started in 2012 by the Open Linguistics group of the Open Knowledge Foundation (McCrae et al.,

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

[2] https://linguistic-lod.org/llod-cloud

2016). The aim was to break the data silos of linguistic data and thus encourage NLP applications that can use data from multiple languages, modalities (e.g., lexicon, corpora, etc.). Looking at the current state of the LLOD, displayed in Figure 1, one can see that the data sets published in the cloud are classified along the lines of six categories:

- Corpora
- Terminologies, Thesauri and Knowledge Bases
- Lexicons and Dictionaries
- Linguistic Resource Metadata
- Linguistic Data Categories
- Typological Databases

Not all the data sets are equally linked to each other, and our project can contribute to better linking of data sets in the fields of Terminologies, Thesauri and Knowledge Bases and those in the fields of Lexicons and Dictionaries.

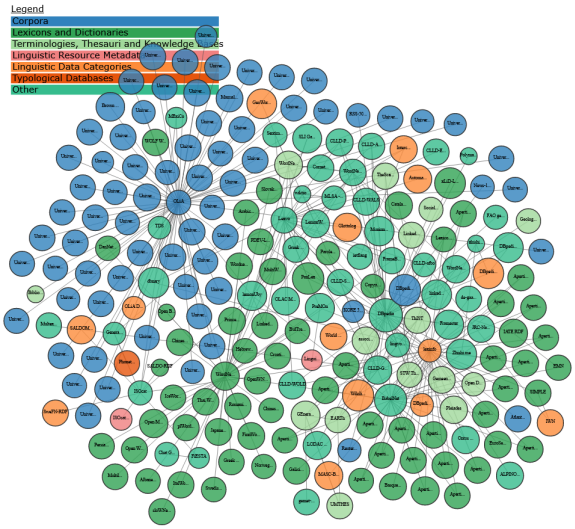

Figure 1: The Linguistic Linked Data Cloud.

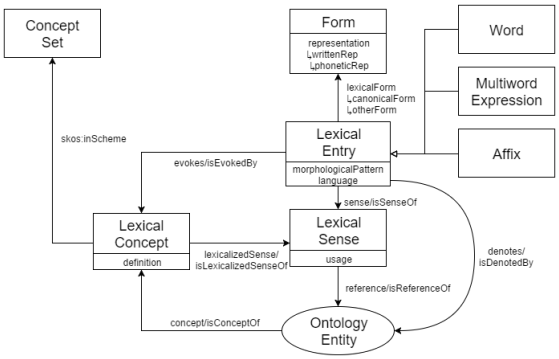

Figure 2: The core Modules of OntoLex-Lemon. Graphic taken from `https://www.w3.org/2016/05/ontolex/`.

3. OntoLex-Lemon

The OntoLex-Lemon model, which resulted from the work of a W3C Community Group,[3] was originally developed with the aim to provide a rich linguistic grounding for ontologies, meaning that the natural language expressions used in the labels, definitions or comments of ontology elements are equipped with an extensive linguistic description.[4] This rich linguistic grounding includes the representation of morphological and syntactic properties of lexical entries as well as the description of the syntax-semantics interface, i.e. the meaning of these lexical entries with respect to an ontology or conceptual scheme.

The main organizing unit for those linguistic descriptions is the *LexicalEntry* class, which enables the representation of morphological patterns for each entry (a multi word expression, a word or an affix). The connection of a lexical entry to an ontological entity is marked mainly by the *denotes* property or is mediated by the *LexicalSense* or the *LexicalConcept* classes, as shown in Figure 2, which displays the core module of the model.

OntoLex-Lemon builds on and extends the *lemon* model (Cimiano et al., 2016). A major difference is that OntoLex-Lemon includes an explicit way to encode conceptual hierarchies relying on SKOS[5] standard. As can be seen in Figure 2, lexical entries can be linked via the *ontolex:evokes* property to such SKOS concepts, which can represent WordNet synsets. This structure is paralleling the relation between lexical entries and ontological resources, which is implemented either directly by the *ontolex:reference* property or mediated by the instances of the *ontolex:LexicalSense* class.

Aside from its original area of application, OntoLex-Lemon has become a de-facto standard in the field of digital lexicography and is being applied for example in the European infrastructure project ELEXIS (European Lexicographic Infrastructure).[6] The extended scope of the model is also reflected in the development of new modules for the OntoLex-Lemon vocabulary. This includes extensions for lexicography (Bosque-Gil and Gracia, 2019), and emerging specifications for morphology (Klimek et al., 2019) and the representation of frequency, attestation and corpus information (Chiarcos and Ionov, 2019). The morphology module is specifically important for the cross-linguistic applicability of OntoLex-Lemon, as it aims to support languages with a lot of stem internal alternations: By using regular expressions to represent morphology generation rules, it provides implementation-independent means to generate inflected forms from lemma information, that can be subsequently incorporated in conventional morphology frameworks such as XFST (Ranta, 1998) or FOMA (Hulden, 2009). Specifications for phonological processes and mor-

[3] See `https://www.w3.org/2016/05/ontolex/`

[4] See (McCrae et al., 2012) and (Cimiano et al., 2016).

[5] SKOS stands for "Simple Knowledge Organization System". SKOS provides "a model for expressing the basic structure and content of concept schemes such as thesauri, classification schemes, subject heading lists, taxonomies, folksonomies, and other similar types of controlled vocabulary" (`https://www.w3.org/TR/skos-primer/`).

[6] See `http://www.elex.is/` for more detail.

phological and syntactic combinatorial restrictions, like restrictions on compounding and derivation are currently being discussed.

4. Main Objectives of Prêt-à-LLOD

The main goal of the project is to allow for multilingual cross-sectoral data access that supports the rapid development of applications and services to be deployed in multilingual cross-border situations. This is realised by providing data *discovery* tools based on metadata aggregated from multiple sources, methodologies for describing the licenses of data and services, and tools to deduce the possible licenses of a resource produced after a complex pipeline. Related to this is the development of a *transformation* platform that maps data sets to the formats and schemas that can be consumed by the LLOD. Finally, the project is developing an ecosystem to support the development of linked data-aware language technologies, from basic tools such as taggers to full applications such as machine translation systems or chatbots, based on semantic technologies that have been developed for LLOD to provide interoperable pipelines. We apply state-of-the-art semantic *linking* technologies in order to provide semi-automatic integration of language services in the cloud.

The sustainability of language technologies and resources is a major concern. We aim to increase sustainability of resources by providing services as data. Using technologies such as Docker,[7] it is possible to wrap services in portable containers that can be shared as single files. Multiple containers can be combined, such that one container takes the output of another as its input, so that services can be aggregated in multi-service workflows. This supports long-term maintenance through methods such as open source software. Furthermore, we build supporting tools to measure and analyse the validity, maintainability and licensing of the data and services. This increases the quality and coverage of language resources and technologies by ensuring that services are easier to archive and reuse, and thus remain available for longer.

5. Methods

This section discusses the methods that are implemented in the project for *discovering*, *transforming* and *linking* linguistic data so that they can be published as LLOD.

5.1. Discovery

Prêt-à-LLOD provides a flexible discovery method that can search over both language resources and services. As many real challenges can only be handled by a combination of multiple datasets and services, the project develops a new workflow system that supports chaining of multiple services using semantic service descriptions and containerization to avoid becoming a "walled garden" ecosystem. A key challenge for this is the chaining of services and data from heterogeneous sources. To this end, we apply linking to develop a transformation component which uses a novel three-step process whereby data from multiple sources is combined by means of RDF (Resource Description Framework,[8] the representation language needed to publish data in the LLOD), linked and then harmonized using semantic and language technologies. The resulting discovery and search platform consist in a single and user-friendly portal. This platform is built on top of the Linghub[9] platfrom, which has been now ported to the CKAN open source data portal platform, ensuring sustainability and scalability.[10]

5.2. Transformation

Existing language data and services operate on different formats. In order to (re)use and combine them in an innovative way, structural and conceptual differences must be overcome.

Prêt-à-LLOD tackles this issue by means of an integrated methodology that transforms language resources. The target models of the transformation are OntoLex-Lemon (briefly introduced above) for lexical data or any RDF vocabulary supporting the representation of language data.

Prêt-à-LLOD integrates many components for the transformation, enrichment and manipulation of language resources within the Flexible Integrated Transformation and Annotation eNgineering (Fintan) platform (Fäth et al., 2020). Fintan is originally based on CoNLL-RDF (Chiarcos and Fäth, 2017), a library developed for round-tripping between tab-separated values and tabular data on the one hand and RDF graphs on the other hand, and for the scalable manipulation of such data by means of parallelized SPARQL Update operations. CoNLL-RDF provides optimizations for popular one-word-per-line formats as conventionally used to represent corpora and linguistic annotations in the language resource community, e.g., in the CoNLL series of shared tasks. Beyond the transformation of corpus data, Fintan has been extended for transforming lexical data sets into RDF representations using OntoLex-Lemon. Fintan currently supports 16 different one-word-per-line corpus formats (CoNLL annotations and variations thereof), 3 one-word-per-line corpus formats with XML extensions (SketchEngine, CWB, TreeTagger), 4 one-word-per-line lexical formats (Unimorph, OpenMultilingual Wordnet, BingLiu sentiment lexicon, TIAD-TSV), and 2 RDF vocabularies (CoNLL-RDF, OntoLex-Lemon).

The project has also implemented transformers that are not yet integrated in Fintan, e.g. Apertium-XML to OntoLex-Lemon, TEI/Dict (FreeDict) to OntoLex-Lemon, UniMorph to OntoLex-Lemon, TBX to OntoLex-Lemon, and PanLex CSV to OntoLex-Lemon.

Selected data sets transformed within the project include:

- RDF conversion of full Apertium data: 55 bilingual dictionaries.
- RDF conversion of the PanLex database: 2,500 dictionaries, compiled into 1,651 substantial bilingual dic-

[7] See `https://hub.docker.com/`.

[8] `https://www.w3.org/RDF/`.

[9] See `http://linghub.org/` and (McCrae and Cimiano, 2015)

[10] The new platform will be soon openly accessible under `https://pret.staging.derilinx.com/`. It is currently (02.03.2020) accessible only with a password, for testing purposes.

tionaries (i.e., those with more than 10,000 entries per language pair).

- RDF conversion of other dictionary collections: 252 bilingual dictionaries (FreeDict, XDXF)
- RDF conversion of morpheme inventories: 110 monolingual morpheme inventories from UniMorph and 7 large-scale morphological resources for 7 EU languages.
- RDF conversion of WordNet(s): three WordNets for Romance languages[11] and one for German.[12]
- Conversion of 5 terminological resources in TBX to RDF.

5.3. Linking

Finally, the project is developing (semi-)automated linking mechanisms. This concerns both the conceptual level of language descriptions as also the lexical data. We are working both in a mono- and in a cross-lingual set up. Since this work is still in progress, at this time we can only report on preliminary approaches.

In the context of cross-lingual concept matching, we are updating an already existent ontology matching tool, CIDER-CL (Gracia and Asooja, 2013) with contemporary techniques based on cross-lingual word embeddings (Artetxe et al., 2018). Regarding linking cross-lingual lexical data, the project has laid the groundwork for research in the topic of "translation inference across dictionaries" by organising the TIAD'19 shared task (Gracia et al., 2019), in which a benchmark and evaluation framework were provided to allow for systematic comparisons between systems. Such systems were able to infer indirect translations between language pairs that were initially disconnected in the Apertium RDF graph (Gracia et al., 2018), showing promising results but also the need of further research.

Ontology lexicalisation aims at developing techniques that can connect existing ontologies to lexicons at larger scale. So far, we are following an unsupervised approach for finding lexicalization patterns via Frequent Subgraph Mining.

The other linking exercises are done with the support of "Naisc" (McCrae and Buitelaar, 2018), a tool developed at the National University of Ireland in Galway, and which is in use within the European Lexicographic Infrastructure (ELEXIS) project.[13]

5.4. Policy-driven Language Resource Discovery and Access

As stated above, Prêt-à-LLOD is also concerned with the issue of detecting and "chaining" licensing conditions for language resources and services, which can be combined in complex pipelines. In addition to the three main contributions concerned with delivery, transformation and linking, the project also deals with the automated execution of smart policies for language data transactions. In particular, part of this work is based on the ODRL specifications.[14]

Since all those steps need to be carefully designed and integrated in a workflow, Prêt-à-LLOD is designing a protocol, based on semantic mark-up, that aims at enabling language services to be easily connected into multi-server workflows.

6. Industry Use Case Pilots

Prêt-à-LLOD involves four industry-led pilot projects that are designed to demonstrate the relevance, transferability and applicability of the methods and techniques under development in the project (cf. Section 5. above) to practical problems in the language technology industry. While Prêt-à-LLOD workflows and methodologies cut across many potential application domains and sectors, the pilots showcase potentials in the context of the following sectors specifically: technology companies, open government services, pharmaceutical industry, and finance. As overarching challenges, all pilots are addressing facets of *cross-language transfer* or *domain adaptation* to varying degrees. In the subsections below, we provide details for each of the pilot projects.

6.1. Multilingual Knowledge Graphs for Knowledge Management across Sectors

In Pilot I, **Semantic Web Company**[15] aims at improving term extraction and concept matching services as offered by their flagship product, PoolParty[16]. While PoolParty is designed to be applicable to multiple domains, the goal is to extend and/or improve terminology extraction and concept matching workflows currently existing in PoolParty with modern open source language resources. In three subpilots, particular attention will be paid to aspects of quality enhancement in term extraction, improved lemmatization capabilities in concept matching, improved word sense disambiguation (WSD) capabilities as a prerequisite for concept matching, and extension to several new languages.

Activities in Pilot I will benefit from Prêt-à-LLOD methodologies in terms of resource discovery, transformation (for the purpose of preparing suitable WSD data sets, in particular) and workflow composition for term and concept extraction approaches. Moreover, Pilot I aims at replacing certain proprietary language resources currently used in PoolParty workflows with open source language resources. For this purpose, tests will be conducted to estimate the loss (or gain) in quality of performance.

6.2. Linking Lexical Knowledge to Facilitate Rapid Integration and Wider Application of Lexicographic Resources for Technology Companies

Oxford University Press,[17] as a provider of highly curated, comprehensive and lexically rich resources for the language technology industry are devoting their activities

[11] See (Racioppa and Declerck, 2019).

[12] See (Declerck et al., 2019).

[13] `https://elex.is/`. See also (McCrae et al., 2019)

[14] ODRL stands for "Open Digital Rights Language" and is a W3C specification (see `https://www.w3.org/TR/odrl-model/`).

[15] `https://semantic-web.com`

[16] `https://www.poolparty.biz`

[17] `https://global.oup.com/`

in Pilot II to explore methodologies for linking lexical data for language services, thus directly addressing one of the key challenges addressed by Prêt-à-LLOD. Two instantiations of the linking challenge will be pursued in respective sub-pilots, viz., linking different dictionaries (either mono- or bilingual ones) at the level of meaning (i.e., sense level), and linking corpus data to dictionary senses by means of word sense disambiguation.

The former follows up the successful linking of key lexical databases, such as WordNet or Wikipedia for language technology purposes (Gurevych et al., 2016; Navigli and Ponzetto, 2012). This has been an area of significant activity in the past decade which more recently has turned into the alignment of dictionary content due to the benefits that dictionary sense linking can contribute (Gracia et al., 2019; McCrae et al., 2017).

With respect to the latter challenge, linking corpora to dictionary senses, Pilots I and II share mutual goals but differ in their conceptual underpinnings. While Pilot II resides in well-established, manually curated sense inventories, Pilot I has an additional focus on word sense induction methods in order to induce sense information on the fly for a given corpus (which might even be applied to specialized domains for which no dictionaries with sense-level information currently exist).

This network of interlinked senses and textual data will open up the possibility of expanding existing lexicographic content with additional data from other sources, and will also enable to port existing lexical resources to new languages (as another commonality with Pilot I).

6.3. Supporting the Development of Public Services in Open Government both within and across borders

In Pilot III, **Derilinx**[18] aims at providing tools and interfaces for intuitive and cross-border access to open data portals using natural language. In the first of two sub-pilots, a web application providing responses to queries regarding public services information via a dashboard will be developed; this involves mapping user queries from natural language into formal queries against an open data health service portal. In a second sub-pilot, the aforementioned application will be developed further into a chatbot providing conversational responses to queries regarding public services information. In both scenarios, cross-language transfer techniques as well as domain adaptation methods based on existing services previously developed at Derilinx are be applied in order to provide a web interface for users to access cross-border open data portals in their native language. To these ends, Derilinx' pilot activities will benefit from all Prêt-à-LLOD services and methodologies, ranging from resource discovery over data management, transformation and linking until workflow composition.

6.4. Multilingual Text Analytics for Extracting Real-World Evidence in the Pharma Sector

In Pilot IV, **Semalytix**[19] focuses on cross-lingual transfer of various types of machine learning models and knowledge resources in order to add multilingual capabilities to their text analytics solutions for generating real-world evidence for customers from the pharmaceutical industry. Real-world evidence is evidence for the effectiveness and safety of a drug product, gathered outside of the controlled settings of clinical trials, in order to demonstrate added value of a drug in terms of improvements in quality of life in specific patient populations. Extracting real-world evidence requires to analyze large volumes of heterogeneous content, including subjective assessments of patients and medical experts, which is typically available as unstructured text in multiple languages.

In developing domain-specific multilingual text analytics services and applications that support real-world evidence generation, the interplay between LLOD resources and (deep) machine learning architectures bears strong potential as enablers of cross-lingual transfer and domain adaptation (Hartung et al., 2020). Prêt-à-LLOD workflows and methodologies will be adopted by Semalytix in order to transform and combine existing LLOD resources, but also to define exhaustive training and evaluation pipelines for language transfer and domain adaptation based on all available LLOD resources satisfying certain criteria.

7. Relations to other Infrastructures

There is a close cooperation and technological interaction with the European Lexicographic Infrastructure (ELEXIS) project.[20] ELEXIS is not only relevant due to the Lexicography use case of Prêt-à-LLOD (led by the partner Oxford University Press (OUP)), but also due to the fact that Linked Data are playing an increasing role in eLexicography, and as such the OntoLex-Lemon model, which is at the core of Prêt-à-LLOD, is getting more and more used in this context of eLexicography (McCrae et al., 2019; Bosque-Gil et al., 2019; Bosque-Gil et al., 2016; Declerck et al., 2017; Stolk, 2019). This connection to ELEXIS is important for Prêt-à-LLOD, as an increasing community of lexicographers are making use of OntoLex-Lemon and other LLOD technologies, thus ensuring sustainability of methods developed within Prêt-LLOD, and also giving feedback on certain issues.

The cooperation with ELEXIS is also concerned with the advancement an increase of compatibility of standards, for example establishing bridges between OntoLex-Lemon, as the result of a W3C Community Group, and the TEI Lex-0 encoding guidelines (Romary and Tasovac, 2018) in development within the TEI community. It is important here to specify which (de-facto) standard is best suited for which aspect of digital lexicography.

Another important relation has been established with the European Language Grid (ELG) project,[21] which has started at the same time as Prêt-à-LLOD, in January 2019. The relation to and cooperation with ELG consists at deploying LLOD services developed by Prêt-à-LLOD in the ELG platform.

A first and successful test has been implemented via containerization of the complete Prêt-à-LLOD service that is

[18] https://derilinx.com

[19] https://www.semalytix.com/

[20] Again, see https://elex.is.

[21] See https://www.european-language-grid.eu/

performing a transformation from a TBX[22] data set into RDF based on OntoLex-Lemon. This is an important achievement as it supports the sustainability of results of the Prêt-à-LLOD project, allowing to deploy its data and services on various platforms besides its core LLOD infrastructure.

Finally, we mention the influential role that Prêt-à-LLOD has played in the newly created European network for Web-centred linguistic data science (NexusLinguarum)[23], a COST Action aimed at promoting synergies across Europe between linguists, computer scientists, terminologists, and other stakeholders in industry and society, in order to investigate and extend the area of linguistic data science. In NexusLinguarum, LLOD technologies will play a central role and the outcomes of Prêt-à-LLOD will be essential to build the holistic ecosystem of multilingual and semantically interoperable linguistic data that NexusLinguarum pursues.

8. Standards

Prêt-à-LLOD members are actively involved in the development, documentation and refinement of community standards and best practices. One recent result of these activities is, among others, the monograph Cimiano et al. (2020) that describes and defines the state of the art in linguistic linked open data.

Beyond that, Prêt-à-LLOD and its members actively participate to on-going standardisation activities, in particular those related to the de-facto standard OntoLex-Lemon, which was published in May 2016 as a report of the W3C Community Group "Ontology-Lexica".[24] As mentioned already, the OntoLex vocabulary has gained an increasing popularity as a means of publishing lexical resources with RDF and as Linked Data. There have been some desiderata to extend the model by new modules to cover for example the representation of (retro-digitized) dictionaries or other linguistic resource containing lexicographic data, and to address structures and annotations commonly found in lexicography. For this, a new module, the "lexicog" module has been developed and published in September 2019.[25] The "lexicog" specifications are a joint product of, among others, colleagues working in ELEXIS and in Prêt-à-LLOD.

While "lexicog" reflects the increasing importance of OntoLex-Lemon for digital lexicography, it does not cover all aspects of digital lexicography, as for example the information that can be derived from corpora: frequency information, links to attestations in corpora, or collocation data, etc. Therefore, the OntoLex community has put forward the proposal for a new module covering frequency, attestation and corpus information (FrAC) that not only covers the requirements of digital lexicography, but also accommodates essential data structures for lexical information in natural language processing. The specifications for this module are already in an advanced stadium.[26] Finally, we would like to mention the development of a morphology module,[27], which is close to its publication. This module supports the encoding of more precise morphological phenomena within OntoLex-Lemon.

9. Conclusion

We presented the current state of the Prêt-à-LLOD project, which is aiming at further extending the Linguistic Linked Open Data cloud infrastructure and making LLOD-compliant services and datasets sustainable. We think of Linguistic Linked Open Data technology and resources as a means to develop a sustainable ecosystem of interoperable, web-based language technology services and language resources in accordance with the early vision formulated by Chiarcos et al. (2013). With the Prêt-à-LLOD project and related infrastructure initiatives, the synergies expected from the use of Linked Data will come within reach within the next years.

As for details and updates on Prêt-à-LLOD tools and resources, please consult our website under `https://www.pret-a-llod.eu/software-and-resource-descriptions/`.

10. Acknowledgements

The project Prêt-à-LLOD has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement no. 825182. This paper is partially based upon work from COST Action CA18209 - NexusLinguarum "European network for Web-centred linguistic data science", supported by COST (European Cooperation in Science and Technology). We also thank the anonymous reviewers for their helpful comments.

11. References

Artetxe, M., Labaka, G., and Agirre, E. (2018). A robust self-learning method for fully unsupervised cross-lingual mappings of word embeddings. In *Proc. of the 56th Annual Meeting ofthe Association for Computational Linguistics*, pages 789–798. Association for Computational Linguistics.

Bosque-Gil, J. and Gracia, J. (2019). The OntoLex Lemon lexicography module. Technical report, W3C Community Group Ontology-Lexica. Final Community Group Report.

Bosque-Gil, J., Gracia, J., and Gómez-Pérez, A. (2016). Linked data in lexicography. *Kernerman Dictionary News*, pages 19–24, jul.

Bosque-Gil, J., Lonke, D., Gracia, J., and Kernerman, I. (2019). Validating the OntoLex-lemon lexicography module with K Dictionaries' multilingual data. In *Electronic lexicography in the 21st century. Proceedings of the eLex 2019 conference.*, pages 726–746, Brno, Czech Republic, October. Lexical Computing CZ s.r.o.,.

[22] TBX stands for "TermBase eXchange". It is an ISO standard (30042:2019) for the representation of terminological data

[23] `https://www.cost.eu/actions/CA18209/`

[24] As a reminder, see `https://www.w3.org/2016/05/ontolex/`.

[25] See `https://www.w3.org/2019/09/lexicog/` and Bosque-Gil et al. (2019).

[26] See `https://github.com/acoli-repo/ontolex-frac`.

[27] `https://www.w3.org/community/ontolex/wiki/Morphology`.

Chiarcos, C. and Fäth, C. (2017). CoNLL-RDF: Linked Corpora Done in an NLP-Friendly Way. In Jorge Gracia, et al., editors, *Language, Data, and Knowledge - First International Conference, LDK 2017, Galway, Ireland, June 19-20, 2017, Proceedings*, volume 10318 of *Lecture Notes in Computer Science*, pages 74–88. Springer.

Chiarcos, C. and Ionov, M. (2019). The OntoLex Lemon module for frequency, attestation and corpus information. Technical report, W3C Community Group Ontology-Lexica. draft version, Mar 3, 2019.

Chiarcos, C., McCrae, J., Cimiano, P., and Fellbaum, C. (2013). Towards open data for linguistics: Linguistic linked data. In Alessandro Oltramari, et al., editors, *New Trends of Research in Ontologies and Lexical Resources: Ideas, Projects, Systems*, pages 7–25. Springer Berlin Heidelberg, Berlin, Heidelberg.

Cimiano, P., McCrae, J. P., and Buitelaar, P. (2016). Lexicon Model for Ontologies: Community Report. W3C Community Group Final Report, World Wide Web Consortium.

Cimiano, P., Chiarcos, C., McCrae, J. P., and Gracia, J. (2020). *Linguistic Linked Data: Representation, Generation and Applications*. Springer International Publishing.

Declerck, T., Tiberius, C., and Wandl-Vogt, E. (2017). Encoding Lexicographic Data in Ontolex: Lessons Learned and Open Questions. In John P. McCrae, et al., editors, *Proceedings of the LDK 2017 Workshops: 1st Workshop on the OntoLex Model (OntoLex-2017), Shared Task on Translation Inference Across Dictionaries & Challenges for Wordnets co-located with 1st Conference on Language, Data and Knowledge (LDK 2017), Galway, Ireland, June 18, 2017*, volume 1899 of *CEUR Workshop Proceedings*, pages 1–10. CEUR-WS.org.

Declerck, T., Siegel, M., and Gromann, D. (2019). Ontolex-lemon as a possible bridge between wordnets and full lexical descriptions. In Christiane Fellbaum, et al., editors, *Proceedings of the Tenth Global Wordnet Conference*, pages 264–271, wyb. Stanisława Wyspiańskiego 27 50-370 Wrocław Poland, 7. Oficyna Wydawnicza Politechniki Wrocławskiej, Oficyna Wydawnicza Politechniki Wrocławskiej.

Fäth, C., Chiarcos, C., Ebbrecht, B., and Ionov, M. (2020). Fintan - Flexible, Integrated Transformation and Annotation eNgineering. In *Proceedings of the Twelfth International Conference on Language Resources and Evaluation (LREC 2020)*, Marseille, France, May 11-16, 2020. European Language Resources Association (ELRA).

Gracia, J. and Asooja, K. (2013). Monolingual and cross-lingual ontology matching with CIDER-CL: Evaluation report for OAEI 2013. In *Proc. of 8th Ontology Matching Workshop (OM'13), at 12th International Semantic Web Conference (ISWC'13)*, volume 1111, Syndey (Australia), oct. CEUR-WS, ISSN-1613-0073.

Gracia, J., Villegas, M., Gómez-Pérez, A., and Bel, N. (2018). The apertium bilingual dictionaries on the web of data. *Semantic Web*, 9(2):231–240, jan.

Jorge Gracia, et al., editors. (2019). *Proceedings of TIAD-2019 Shared Task - Translation Inference Across Dictionaries co-located with the 2nd Language, Data and Knowledge Conference (LDK 2019), Leipzig, Germany, May 20, 2019*, volume 2493 of *CEUR Workshop Proceedings*. CEUR-WS.org.

Gurevych, I., Eckle-Kohler, J., and Matuschek, M. (2016). *Linked Lexical Knowledge Bases: Foundations and Applications*. Synthesis Lectures on Human Language Technologies. Morgan & Claypool Publishers.

Hartung, M., Orlikowski, M., and Veríssimo, S. (2020). Evaluating the impact of bilingual lexical resources on cross-lingual sentiment projection in the pharmaceutical domain. http://doi.org/10.5281/zenodo.3707940.

Hulden, M. (2009). Foma: a finite-state compiler and library. In *Proceedings of the Demonstrations Session at EACL 2009*, pages 29–32, Athens, Greece, April. Association for Computational Linguistics.

Klimek, B., McCrae, J. P., Bosque-Gil, J., Ionov, M., Tauber, J. K., and Chiarcos, C. (2019). Challenges for the representation of morphology in ontology lexicons. In *Proceedings of eLex 2019. Electronic lexicography in the 21st century: Smart lexicography*.

McCrae, J. P. and Buitelaar, P. (2018). Linking Datasets Using Semantic Textual Similarity. *Cybernetics and Information Technologies*, 18(1):109–123.

McCrae, J. P. and Cimiano, P. (2015). Linghub: a linked data based portal supporting the discovery of language resources. In Agata Filipowska, et al., editors, *Joint Proceedings of the Posters and Demos Track of 11th International Conference on Semantic Systems - SEMANTiCS 2015 and 1st Workshop on Data Science: Methods, Technology and Applications (DSci15) 11th International Conference on Semantic Systems - SEMANTiCS 2015, Vienna, Austria, September 15-17, 2015*, volume 1481 of *CEUR Workshop Proceedings*, pages 88–91. CEUR-WS.org.

McCrae, J., de Cea, G. A., Buitelaar, P., Cimiano, P., Declerck, T., Gómez-Pérez, A., Gracia, J., Hollink, L., Montiel-Ponsoda, E., Spohr, D., and Wunner, T. (2012). Interchanging lexical resources on the Semantic Web. *Language Resources and Evaluation*, 46(6):701–709.

McCrae, J. P., Chiarcos, C., Bond, F., Cimiano, P., Declerck, T., de Melo, G., Gracia, J., Hellmann, S., Klimek, B., Moran, S., Osenova, P., Pareja-Lora, A., and Pool, J. (2016). The Open Linguistics Working Group: Developing the Linguistic Linked Open Data Cloud. In Nicoletta Calzolari (Conference Chair), et al., editors, *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16)*, 9, rue des Cordelières, 75013 Paris, 5. ELRA, ELRA.

John P. McCrae, et al., editors. (2017). *Proceedings of the LDK 2017 Workshops: 1st Workshop on the OntoLex Model (OntoLex-2017), Shared Task on Translation Inference Across Dictionaries & Challenges for Wordnets co-located with 1st Conference on Language, Data and Knowledge (LDK 2017), Galway, Ireland, June 18, 2017*, volume 1899 of *CEUR Workshop Proceedings*. CEUR-WS.org.

McCrae, J. P., Tiberius, C., Khan, A. F., Kernerman, I., Declerck, T., Krek, S., Monachini, M., and Ahmadi, S.

(2019). The ELEXIS Interface for Interoperable Lexical Resources. In Iztok Kosem, et al., editors, *Proceedings of the eLex 2019 conference. Biennial Conference on Electronic Lexicography (eLex-2019), Electronic lexicography in the 21st century, October 1-3, Sintra, Portugal*, pages 642–659. CELGA-ILTEC, University of Coimbra, Lexical Computing CZ, s.r.o, 10.

Navigli, R. and Ponzetto, S. P. (2012). BabelNet: The automatic construction, evaluation and application of a wide-coverage multilingual semantic network. *Artif. Intell.*, 193:217–250.

Racioppa, S. and Declerck, T. (2019). Enriching open multilingual wordnets with morphological features. In Raffaella Bernardi, et al., editors, *Proceedings of the Sixth Italian Conference on Computational Linguistics*. CEUR, 10.

Ranta, A. (1998). A multilingual natural-language interface to regular expressions. In Lauri Karttunen et al., editors, *Proceedings of the International Workshop on Finite State Methods in Natural Language Processing (FSMNLP'98)*, Ankara, Turkey.

Romary, L. and Tasovac, T. (2018). TEI Lex-0: A Target Format for TEI-Encoded Dictionaries and Lexical Resources. In *TEI Conference and Members' Meeting*, Tokyo, Japan, September.

Stolk, S. (2019). A Thesaurus of Old English as Linguistic Linked Data: Using OntoLex, SKOS and lemon tree to Bring Topical Thesauri to the Semantic Web. In Iztok Kosem, et al., editors, *Proceedings of the eLex 2019 conference. Biennial Conference on Electronic Lexicography (eLex-2019), Electronic lexicography in the 21st century, October 1-3, Sintra, Portugal*, pages 223–247. CELGA-ILTEC, University of Coimbra, Lexical Computing CZ, s.r.o, 10.