Linguistic Linked Open Data for All

John McCrae[1], Thierry Declerck[2]
[1]Insight Centre for Data Analytics at the National University of Ireland Galway, Ireland
[2]DFKI GmbH, Multilinguality and Language Technology, Germany
[1] john.mccrae@insight-centre.org
[2]declerck@dfki.de

Abstract

In this paper we briefly describe the European H2020 project "Prêt-à-LLOD" ('Ready-to-use Multilingual Linked Language Data for Knowledge Services across Sectors'). This project aims to increase 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. Prêt-à-LLOD aims to achieve 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, in particular the usage of the LLOD.

Keywords: Linguistic Linked Open Data, Standards, Infrastructure

Résumé

Dans cet article, nous décrivons brièvement le projet Européen «Prêt-à-LLOD» («Données multilingues prêt à l'emploi pour les services de la connaissance dans tous les secteurs»). Ce projet vise à accroître l'utilisation des technologies langagières en exploitant la combinaison de données liées et de technologies langagières, à savoir les données linguistiques ouvertes et liées (LLOD), pour créer des données multilingues prêtes à l'emploi. Prêt-à-LLOD vise à atteindre cet objectif en créant une nouvelle méthodologie pour construire des chaînes de valeur de données applicables à un large éventail de secteurs et d'applications et reposant sur des ressources linguistiques et des technologies langagières pouvant être intégrées au moyen de technologies sémantiques, en particulier l'utilisation du LLOD

1. Introduction

Language technologies increasingly rely on large amounts of data and better access and usage of language resources will enable to provide multilingual solutions that support the further development of language technologies in Europe and in the world. 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. An essential part of the Extract-Transform-Load process currently needed involves linking datasets to existing schemas, yet few specialists take advantage of linked data technologies to perform this task. 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. Prêt-à-LLOD aims to achieve 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 novel tools for the discovery, transformation and linking of datasets and apply these to both data and metadata in order to provide multi-portal access to heterogeneous data repositories. A goal is also 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 deployed to many platforms, including as-yet-unknown platforms, and can be self-described with linked data semantics. Our approach is being validated in four pilots.

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 on. We then discuss briefly some of the objectives of the project methodologies put in place in order to reach them, showing also their relevance for less-resourced languages.

2. Linguistic Linked Open Data Cloud

The Linguistic Linked Open Data (LLOD) cloud[2] is an initiative, which was started in 2012 by a group of the Open Knowledge Foundation[3]. 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.) and develop novel algorithms. Looking at the current state of the LLOD, displayed in Figure 1, one can see that the data sets published in this cloud are classified along the lines of six categories:

- Corpora
- Terminologies, Thesauri and Knowledge Bases
- Lexicons and Dictionaries

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

[2] See https://linguistic-lod.org/llod-cloud for more detail.

[3] See (McCrae et al., 2016) for a description of the development of the LLOD.

- Linguistic Resource Metadata
- Linguistic Data Categories
- Typological Databases

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

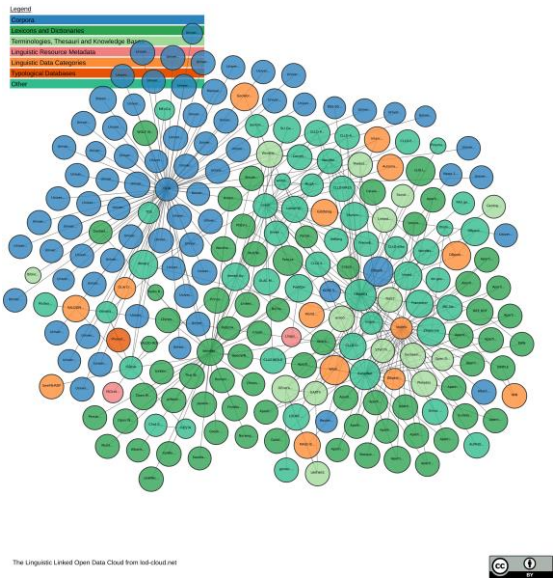

Figure 1: The Linguistic Linked Data Cloud

3. OntoLex-Lemon

The OntoLex-Lemon model, which is resulting from a W3C Community Group[4], 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.[5] This rich linguistic grounding includes the representation of morphological and syntactic properties of lexical entries as well as the syntax-semantics interface, i.e. the meaning of these lexical entres with respect to an ontology or to specialized vocabularies.

The main organizing unit for those linguistic descriptions is the lexical entry, 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 this is represented 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, using the SKOS[6] 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.

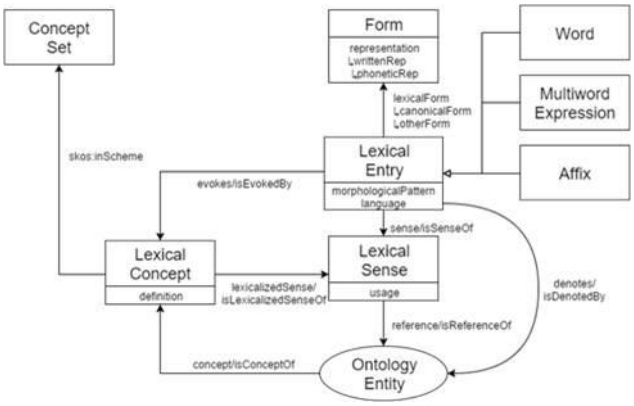

Figure 2: The core Modules of OntoLex-Lemon.
Graphic taken from https://www.w3.org/2016/05/ontolex/.

More recently, OntoLex-Lemon has been used also as 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)[7].

4. Main Objectives of Prêt-à-LLOD

The first goal of this 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.

A second goal consists in developing a new ecosystem to support the development of novel 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

A third goal is concerning sustainability. The sustainability of language technologies and resources is a major concern. We aim to solve this by providing services as data , that is, wrapping services in portable containers that can be shared as single files. Language data also eventually becomes valueless as the documentation and expertise for processing esoteric formats is lost, and apply the paradigm of data as services, where services can be embedded in multi-service workflows, that demonstrates the service's value and

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

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

[6] 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/).

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

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. In particular, this goal is important for minority and under-resourced languages, where the precious effort to develop resources is often lost.

5. Methods

The project implements methods for discovery, transforming and linking linguistic data so that they can be published in the LLOD.

Prêt-à-LLOD provide 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, 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 consists in a single and user-friendly portal.

An integrated methodology has been designed for the transformation of language resources. The goal of the transformation is either OntoLex-Lemon model (briefly introduced above) for lexical data or any RDF vocabulary supporting the representation of language data. This is an important aspect for less- or under-resourced languages, as they have the same "representation dignity" as other languages, to which they can be linked to, in the LLOD ecosystem.

Finally, the project is developing (semi-)automated linking mechanisms. This concerns both conceptual level of language descriptions as also the lexical data. We are working both in a mono- and in a cross-lingual set up.

As stated above, Prêt-à-LLOD is also concerned with the issue of detecting and "chaining" licensing conditions for the language resources and services that can be combined in complex pipelines. So that additionally to the three basic methodologies described just above, the project is dealing with the automated execution of smart policies for language data transactions.

All those steps need for sure to be carefully designed and integrated in a workflow. Prêt-à-LLOD is therefore designing a protocol, based on semantic markup, that is aiming at enabling language services to be easily connected into multi-server workflows.

6. Standards

Prêt-à-LLOD members are involved in a series of standardisation activities, mainly related to the de-facto standard OntoLex-Lemon. We would like to mention here the new module on lexicography ("lexicog")[8], on a more precise description of morphological phenomena[9] and on the topics of FRequency, Attestations and Corpus data ("frac")[10]. Those new modules are extending the expressive power of OntoLex-Lemon, and also very important for the inclusion of less- and under-resourced language data in the LLOD, as the scope of the de-facto standard is extended to corpus data, besides the coverage of lexical data.

7. 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 more language data interoperable, also with sustainable semantic description approaches.

8. 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.

9. References

Cimiano, Philipp, McCrae, John and Paul Buitelaar. 2016. Lexicon Model for Ontologies: W3C Community Report.

McCrae, John, Aguado-de Cea, Guadalupe, Buitelaar, Paul, Cimiano, Philipp, Declerck, Thierry, Gomez-Perez, Asuncion, Garcia, Jorge, Hollink, Laura, Montiel-Ponsoda, Elena, Spohr, Dennis and Wunner, Tobias. 2012. Interchanging Lexical Resources on the Semantic Web. Language Resources and Evaluation, 46(4):701–719

John P. McCrae, Christian Chiarcos, Francis Bond, Philipp Cimiano, Thierry Declerck, Gerard de Melo, Jorge Gracia, Sebastian Hellmann, Bettina Klimek, Steven Moran, Petya Osenova, Antonio Pareja-Lora, Jonathan Pool. The Open LinguisticsWorking Group: Developing the Linguistic Linked Open Data Cloud Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16), Portoroˌ, Slovenia, ELRA, ELRA, 9, rue des Cordelières, 75013 Paris, 5/2016

Rodiguez-Doncel,Victor, Casanovas, Pompeu. 2018. A Linked Data Terminology for Copy-right Based on Ontolex-Lemon: AICOL. International Workshops 2015-2017. 410-423.

[8] See https://www.w3.org/2019/09/lexicog/

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

[10] https://github.com/acoli-repo/ontolex-frac