Project abbreviation: Prêt-à-LLOD

Project name: Ready-to-use Multilingual Linked Language Data for Knowledge Services across Sectors

Project coordinator: National University of Ireland Galway (Ireland)

Project consortium:

- National University of Ireland Galway (Ireland)
- Universidad Politécnica de Madrid (Spain)
- Universidad de Zaragoza (Spain)
- Goethe Universität Frankfurt (Germany)
- University of Bielefeld (Germany)
- DFKI (Germany)
- Semantic Web Company (Austria)
- Semalytix (Germany)
- Oxford University Press (United Kingdom)
- Derilinx (Ireland)

Funding: H2020-EU.2.1.1 -- € 2 997 181,25

Project duration: 1 January 2019 - 30 June 2022 (42 months)

Main key words: linguistic linked data, language resource sustainability, Web services for natural language processing, Semantic Web Technologies, Multilingualism

Goal of the project: Language technologies that rely on large amounts of data and better access to language resources permit the delivery of multilingual solutions to support Europe's Digital Single Market. However, language technology specialists spend 80 % of their time cleaning, organising and collecting data sets because data is not 'ready-to-use'. Although an essential part of the extract-transform-load process requires linking data sets to existing designs, linked data technologies remain unexploited. Prêt-a-LLOD increases the use of language technologies to create ready-to-use multilingual data. The project will combine linked data sets with language technologies that are Linguistic Linked Open Data (LLOD) and develop innovative tools for the transformation and linking of data sets.

Project abstract: Language technologies increasingly rely on large amounts of data and better access and usage of language resources will enable to provide multilingual solutions that would support the emerging Digital Single Market in Europe. However, data is rarely 'ready-to-use' and language technology specialists spend over 80% of their time on cleaning, organizing and collecting 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 involves linking datasets to existing schemas, yet few specialists take advantage of linked data technologies to perform this task. In this project we aim 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 Linguistic Linked Open Data (LLOD). The project will develop novel tools for the transformation and linking of datasets and apply these to both data and metadata to provide

multi-portal access to heterogeneous data repositories. We will study how we can automatically analyse licenses to deduce how data may be lawfully used and sold by language resource providers. Finally, we will provide tools to combine language services and resources into complex pipelines by use of semantic technologies. This will lead 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. This toolkit will be validated in four pilots, where novel data value chains will be built for pharmaceutical applications, technology providers, and government services.

Publications:

Recent Developments for the Linguistic Linked Open Data Infrastructure (LREC 2020)

Defying Wikidata: Validation of Terminological Relations in the Web of Data (LREC 2020)

Leveraging Linguistic Linked Data for Cross-Lingual Model Transfer in the Pharmaceutical Domain (ISWC 2021)

Linguistic Linked Data (book published by Springer)

Proceedings of the 7th Workshop on Linked Data in Linguistics (LDL-2020)