



D5.4 Community Building Report

Author(s): Thierry Declerck (DFKI),
John McCrae (NUIG), Jorge Gracia
(UNIZAR)

Date: 31.12.2019



H2020-ICT-29b

Grant Agreement No. 825182

Prêt-à-LLOD - Ready-to-use Multilingual
Linked Language Data for Knowledge
Services across Sectors

D5.4 Community Building Report

Deliverable Number: D5.4

Dissemination Level: PU

Delivery Date: 31.12.2019

Version: Final

Author(s): Thierry Declerck (DFKI), John McCrae (NUIG), Jorge Gracia (UNIZAR)

Document History

Version Date	Changes	Authors
30.09.2019	First version	Thierry Declerck, John McCrae
21.10.2019	Second version	Thierry Declerck, John McCrae, Jorge Gracia
23.10.2019	Editing comments by partners Derilinx (Katharine Cooney) and NUIG (John P. McCrae)	Thierry Declerck
20.11.2019	Preliminary version submitted to the Commission and to the project reviewers	Thierry Declerck, John McCrae
30.12.2019	Final version	Thierry Declerck



Table of Contents

1 Introduction	5
2 Relations to the other projects resulting from the call “ICT-29-2018: A multilingual Next Generation Internet”	6
3 Training Activities of Prêt-à-LLOD	7
3.1 The 3rd Summer Datathon on Linguistic Linked Open Data (SD-LLOD-19)	7
3.2 Contributions to Spring and Summer Schools	8
4 Academic Conferences and Workshops organized by Prêt-à-LLOD	9
4.1 The Language Data and Knowledge (LDK) Conference organized by Prêt-à-LLOD and co-located Events	9
4.2 The Translation Inference Across Dictionaries (TIAD 2019) shared Task	10
4.3 W3C OntoLex Community Group Meeting	10
4.4 Digital Humanities 2019 Workshop: eLexicography between Digital Humanities and Artificial Intelligence	10
4.5 TEI-OntoLex Workshop	11
5 Relations of Prêt-à-LLOD to the Lexicography Community	11
6 Relations of Prêt-à-LLOD to the WordNet Community	12
7 Relations of Prêt-à-LLOD to the Terminology Community	12
8 Relations of Prêt-à-LLOD to European Initiatives	13
9 Industry Events	14
9.1 The first Prêt-à-LLOD Innovation Event: The ELSE-IF workshop @ SEMANTICS 2019	15
9.2 Prêt-à-LLOD at the BDV PPP Meetup 2019	16
9.3 Prêt-à-LLOD and ELG jointly organized the “Initial Workshop on the Language Technology Market and Components Taxonomy” at the LT-Innovate Summit 2019	16
10 The first Sustainability Event of Prêt-à-LLOD	16
11 The Linked Data for Language Technologies (LD4LT) Community Group Report	17
12 Prêt-à-LLOD Standardisation Activities	18
13 Contributions to the new COST Action CA18209 - European network for Web-centred linguistic data science (“NexusLinguarum”)	18
14 Conclusion	19
15 References	19



D5.4.1 Community Building Report

1 Introduction

This deliverable reports on activities pursued in the context of Task 5.4. “Community Building”, which is a part of WP5 “Language Resource and Service Sustainability”. It also contains a summary of the “Innovation” and “Sustainability” events to be organized by the project once every year.

Task 5.4 deals with raising awareness on the potentials of Linguistic Linked Open Data (LLOD) driven language resources and services, beyond the consortium and related initiatives. It is complementing the dissemination strategy of Prêt-à-LLOD, which is described in Section 2 of the deliverable on the Project Dissemination Report (D6.1.1).

A way to achieve this goal is to contribute to the extension of existing standards, or even in the creation of new ones, if needed. Partners of Prêt-à-LLOD have intensively explored their relations to the relevant ISO and W3C committees. Further, discussions with the Text Encoding Initiative (TEI) community have been initiated. Such activities aim to establish a community extension across standardisation bodies. Prêt-à-LLOD standardisation activities are reported in detail for the first year of the project in the deliverable D5.1 “Report on Vocabularies for Interoperable Language Resources and Services”. From year 2 on they will be reported in the corresponding versions of the deliverable for task 5.4: D5.4.2 and D5.4.3.

A second type of activities consisted in organizing LLOD-specific hackathons or datathons and in contributing to spring and summer schools, complementing thus the T6.1 dissemination activities (see D6.1.1) with in-depth training activities.

A third type of community building efforts consisted of presenting the methodologies and results of Prêt-à-LLOD in industry events, especially those relevant to the WP4 use cases. Concerning the Government use case, Prêt-à-LLOD also contributed to a specific workshop organized in the context of the Connecting European Facilities (CEF) activities, as well as to an event organized by the European Publication Office.

Finally, Prêt-à-LLOD is engaged in cooperation with the other projects resulting from the call “ICT-29-2018: A multilingual Next Generation Internet”.

The following sections will describe in detail the activities of the project reflecting the points made above, as well as the first “Innovation” and “Sustainability” events organized by Prêt-à-LLOD.



2 Relations to the other projects resulting from the call “ICT-29-2018: A multilingual Next Generation Internet”

From the very beginning, Prêt-à-LLOD aimed at establishing strong relations to the other projects funded in the context of the “ICT-29-2018: A Multilingual Next Generation Internet” call. Two types of projects were foreseen in this call: 1) an Innovation Action, called “A European Language Grid”, corresponding to the call “29a”, and 2) a Research and Innovation Action, called “Domain-specific/challenge-oriented Human Language Technology”, corresponding to call “29b”. We describe below briefly the main objectives of the projects that have been selected for this call.

One project was selected for the 29a call and the project has the name “European Language Grid” (ELG: see <https://www.european-language-grid.eu/>, or at the Cordis site: <https://cordis.europa.eu/project/rcn/219378/factsheet/en>). “The ELG will be a scalable cloud platform, providing, in an easy-to-integrate way, access to hundreds of commercial and non-commercial Language Technologies for all European languages, including running tools and services as well as data sets and resources. It will enable the commercial and non-commercial European LT community to deposit and upload their technologies and data sets into the ELG, to deploy them through the grid, and to connect with other resources.” (quoted from <https://cordis.europa.eu/project/rcn/219378/factsheet/en>)

6 projects were selected for the call 29b. They are listed just below.

- Bergamot (Browser-based Multilingual Translation): <https://browser.mt/> or at the Cordis site: <https://cordis.europa.eu/project/rcn/219608/factsheet/en>.
- COMPRISE (Cost-Effective, Multilingual, Privacy-driven, Voice-enabled Services): <https://www.compriseh2020.eu/> or at the Cordis site: <https://cordis.europa.eu/project/rcn/218720/factsheet/en>.
- ELITR (European Live Translator): <https://elitr.eu/> or at the Cordis site: <https://cordis.europa.eu/project/rcn/219124/factsheet/en>.
- EMBEDDIA (Cross-Lingual Embeddings for Less-Represented Languages in European News Media): <http://embeddia.eu/> or at the Cordis site: <https://cordis.europa.eu/project/rcn/218731/factsheet/en>.
- GoURMET (Global Under-Resourced MEedia Translation): <https://gourmet-project.eu/> or at the Cordis site: <https://cordis.europa.eu/project/rcn/221392/factsheet/en>.
- Prêt-à-LLOD (Ready-to-use Multilingual Linked Language Data for Knowledge Services across Sectors): <http://www.pret-a-llod.eu/> or at the Cordis site: <https://cordis.europa.eu/project/rcn/220250/factsheet/en>.

A first “target” for Community Building activities in Prêt-à-LLOD is given by the list of those associated projects, while all “29b” projects are invited to deliver their resulting data and



tools to the European Language Grid infrastructure. Prêt-à-LLOD already delivered a service to ELG, as a testbed for a larger number of LLOD services to be integrated in ELG, in parallel to the LLOD cloud, which is the primary infrastructure to be addressed by the project. This service, called “Terme-à-LLOD” originated within the Pilot IV of Prêt-à-LLOD: “Multilingual Text Analytics for Extracting Real-World Evidence in the Pharma Sector”, and is transforming terminological databases into linguistic linked data representations.

Behind this technological cooperation, the 7 projects are aiming at discovering joint topics to work on. For this, very soon a regular teleconference meeting has been agreed on, and representatives of all projects are joining in such a virtual meeting every six weeks.

One of the topics that Prêt-à-LLOD has been addressing in those teleconferences is the issue dealing with sustainability, aiming at establishing a joint document collecting all the sustainability strategies to be implemented in the various projects. Preliminary results of this initiative has been presented at the Meta-Forum 2019 event¹, within a session dedicated to “Sustainability”. This plenary session was jointly organized by Prêt-à-LLOD and ELG. Prêt-à-LLOD was additionally present at META-FORUM with a plenary presentation and a booth.

Another community building cooperation has been established between Prêt-à-LLOD and ELG, with both projects contributing to the further development of the so-called Meta-Share Ontology (ms-ontology) dealing with the refinement and extension of the current metadata infrastructure used in the context of Meta-Share and to be further developed for ELG and other projects (see <http://www.meta-share.org/p/93/Documentation#model>, Gavrilidou et al. (2012)).

A specific cooperation will be established with the EMBEDDIA project in integrating cross-lingual word embeddings, represented as dense real-valued vectors, with lexical data that are encoded in the OntoLex-Lemon model, based on a symbolic representation of lexical knowledge.

3 Training Activities of Prêt-à-LLOD

3.1 The 3rd Summer Datathon on Linguistic Linked Open Data (SD-LLOD-19)

The Summer Datathon on Linguistic Linked Open Data (SD-LLOD) is a series of hackathons that encourage the use of linguistic linked open data in linguistics. In 2019, Prêt-à-LLOD organized the third edition of this at Schloss Dagstuhl² in Germany after two previous successful instances held in Madrid in 2015 and 2017. The datathon was attended by 47 people, including participants, tutors, organisers, and invited speakers, and was organized by Christian Chiarcos (GUF), John McCrae (NUIG) and Jorge Gracia (UNIZAR). The

¹ <https://www.european-language-grid.eu/meta-forum-2019/>

² More details at <https://datathon2019.linguistic-lod.org/>



datathon consisted of a number of talks, practical sessions and most importantly hands-on hacking sessions, where participants developed linguistic resources as linked open data.

The hacking sessions were divided into 8 mini-projects, each supervised by a tutor. The projects were: 1) Redefining the Dictionary Fabric - the case from industry, 2) Linked Linguistic Corpus Data, 3) TED talks transformation to LOD (was the winner of the best mini-project award), 4) Modeling Cherokee lexical-morphological data, 5) Historical Texts -- Linking multiple annotations: Corpora, 6) Ontology Conversion and Resource Population, 7) Modelling Akkadian dictionary data with ontolox-lemon using lexicog, and 8) e-science corpus metadata as Linked Open Data. The description of those projects will soon be published in a Dagstuhl Publication.

The program had two invited talks by Richard Eckart de Castilho (TU Darmstadt) and Gerard de Melo (Rutgers University) and seminars developed by Prêt-à-LLOD members on LLOD basics, ontologies, metadata, annotations and the OntoLex model.

As with previous instances of the datathon, new datasets and research papers have originated from the datathon, including until now:

- “Creating a Multilingual Terminological Resource using Linked Data: the case of Archaeological Domain in the Italian language.” Giulia Speranza, Carola Carlino, Conferenza italiana sulla linguistica computazionale (CLiC-it 2019).
- “Automatic Creation of Linked Legal Terminologies”. Patricia Martín-Chozas, Sina Ahmadi, Elena Montiel-Ponsoda, submitted to The 35th ACM/SIGAPP Symposium On Applied Computing.

3.2 Contributions to Spring and Summer Schools

Thierry Declerck (DFKI) and John McCrae (NUIG) have been invited to give a lecture at the Spring School LiSeH 2019, Graz, Austria, April 23-27, 2019. The topic of the Spring School was “Linked Data and the Semantic Web for Humanities research”. This invitation allowed Prêt-à-LLOD to present in one session an “Introduction to Linked Open Data in Linguistics” to the Digital Humanities community. 25 participants were attending, all selected by the organization committee on the basis of application forms submitted by the interested people.

³

John McCrae (NUIG) and Thierry Declerck (DFKI) gave a one week course on the topic “Introduction to Linked Open Data in Linguistics” at the 31st edition of ESSLLI (European Summer School in Logic, Language and Information), which took place from 5 August to 16 August 2019 at University of Latvia in Riga, Latvia. The course was divided in 5 units, which are a large extension to the lecture given at LiSeH, and which included sessions on “Language Resources and RDF Vocabularies, including SPARQL”, “An Introduction to the OntoLex-Lemon Model”, “Corpus Annotation with LLOD”, “WordNets and TEI-Lex” and “The

³ More details on the LiSeH Spring School:

<https://informationsmodellierung.uni-graz.at/de/veranstaltungen/spring-school-liseh-2019/>



Linked Open Data Cloud”. The course attracted 30 participants from the broader fields of Computer Science and Computational Linguistics.⁴

John McCrae recorded the sessions, which are now available as a YouTube video: <https://www.youtube.com/playlist?list=PLXhZBBpMXr0K867wGyvPIrfUE6F9oqBVV>

4 Academic Conferences and Workshops organized by Prêt-à-LLOD

4.1 The Language Data and Knowledge (LDK) Conference organized by Prêt-à-LLOD and co-located Events

Language, Data and Knowledge (LDK) is a bi-annual conference series on matters of human language technology, data science, and knowledge representation, initiated in 2017 by a consortium of researchers from two Prêt-à-LLOD partners (NUIG, GUF) and the University of Leipzig. It has been supported by an international Scientific Committee of leading researchers in Natural Language Processing, Linked Data and Semantic Web, Language Resources and Digital Humanities. The second edition of the LDK conference was hosted 2019 by the Institut für Angewandte Informatik (InfAI) in Leipzig (May 20-23), and co-organized by the two Prêt-à-LLOD partners NUIG and GUF.

LDK conferences aim at bringing together researchers from across disciplines concerned with the acquisition, curation and use of language data in the context of data science and knowledge-based applications. The efficient and meaningful exploitation of this data in scientific and commercial innovation is at the core of data science research, employing natural language processing and machine learning methods as well as semantic technologies and knowledge graphs.

As in the previous edition, LDK 2019 featured several collocated satellite events dedicated to the conference topics. This included the 13th DBpedia community meeting, the 2nd Shared Task on Translation Inference Across Dictionaries (TIAD), organized by the project partner UNIZAR, a workshop of the W3C Ontology Lexica Community, organized by Prêt-à-LLOD, and a tutorial on historical text reuse (TRACER).⁵

In total, 43 papers were submitted and reviewed by 88 reviewers (at least 3 reviews per paper), resulting in 26 accepted papers. The proceedings of the main conference were published through Dagstuhl Publishing’s OASICs series (Eskevich et. al., 2019) and are available at the following URL: <http://www.dagstuhl.de/dagpub/978-3-95977-105-4>.

The conference programme additionally encompassed invited talks on Mapping the Lexicons of Signs and Words by Christiane Fellbaum (Princeton University), and on Schema.org Annotations and Web Tables: Underexploited Semantic Nuggets on the Web? by Christian

⁴ More information on ESSLLI 2019 can be found here: <https://esslli2019.folli.info/programme/week-1/>.

⁵ See <http://2019.ldk-conf.org/> for more details.



Bizer (Mannheim University), as well as on The Sorbian languages by Eduard Werner (University of Leipzig).

In the context of LDK 2019, Thierry Declerck (DFKI) and John McCrae (NUIG) organized a poster track, inviting for submissions on all research topics that are related to language, data and knowledge, including, but not restricted to most recent developments, future directions or updates to established approaches within the scope of LDK-2019. We were happy to receive 12 abstracts, each reviewed by 3 reviewers. 9 abstracts were accepted and were included in a standalone CEURS proceedings (<http://ceur-ws.org/Vol-2402/>), in an extended form (see Declerck and McCrae (2019)). The topics addressed by the posters were related to Digital Humanities and Lexicography, with a focus on Linked Data.

4.2 The Translation Inference Across Dictionaries (TIAD 2019) shared Task

Co-located with LDK 2019, the second shared task for Translation Inference Across Dictionaries (TIAD 2019), co-organized by the Prêt-à-LLOD partner UNIZAR (Jorge Gracia and team), took place on May 20. This initiative is aiming at exploring methods and techniques for automatically generating new bilingual (and multilingual) dictionaries from existing ones, by supporting a coherent experiment framework that enables reliable validation of results and solid comparison of the processes used.⁶ Four systems participated in the shared task. The participant teams submitted a system description paper including: a description of their system, the way data was processed, the applied algorithms, the obtained results, as well as the conclusions and ideas for future improvements. An additional paper was submitted as regular paper describing a system that did not participate in the shared task but that described a technique that was relevant for the topic of the workshop. The proceedings of TIAD 2019 include a sixth paper, consisting of a summary in which the TIAD organisers give more details about the task, the evaluation data, and the system results.⁷

4.3 W3C OntoLex Community Group Meeting

Partners of the Prêt-à-LLOD consortium organized an OntoLex workshop, in which two new modules for OntoLex were discussed, the “Morphology”⁸ and the “Frequency, Attestation and Corpus Information”⁹ modules, as well as future activities of the W3C Community Group. This workshop was attended by 25 participants.

⁶ See <http://tiad2019.unizar.es> for more details.

⁷ The TIAD proceedings have been published by CEURS Press: <http://ceur-ws.org/Vol-2493/>

⁸ See <https://www.w3.org/community/ontolex/wiki/Morphology> for more details.

⁹ See <https://acoli-repo.github.io/ontolex-frac/> for more details.



4.4 Digital Humanities 2019 Workshop: eLexicography between Digital Humanities and Artificial Intelligence

In a cooperation between the NUIG partner and ELEXIS a joint workshop on “eLexicography between Digital Humanities and Artificial Intelligence” was organized at the Digital Humanities Conference 2019 in Utrecht (9-12 July 2019), which attracted around 30 participants. There Paul Buitelaar (NUIG) also presented an overview of Prêt-à-LLOD.¹⁰

4.5 TEI-OntoLex Workshop

In collaboration with the ELEXIS H2020 infrastructure project, a two day workshop connecting the main developers of the OntoLex-Lemon model and the TEI Lex-0 model was held on the 4th and 5th of October in Sintra, Portugal after the eLex 2019 conference. In this meeting, discussions were held about the interaction of the two models and their differences. As a principal result, it was surmised that the models were complementary as TEI-Lex0 aims to maintain the dictionary as an artifact, while OntoLex-Lemon is a data model. As a result, researchers from both sides agreed to create an authoritative mapping between TEI and OntoLex as an XSLT stylesheet on GitHub at <https://github.com/elexis-eu/tei2ontolex>.

5 Relations of Prêt-à-LLOD to the Lexicography Community

The relation to the European Lexicographic Infrastructure (ELEXIS) project (<https://elex.is>) is not only relevant due to the Lexicography use case of Prêt-à-LLOD (led by the partner Oxford University Press (OUP)), but also 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 (McCrae et al., 2019; Bosque-Gil et al., 2019; Declerck et al., 2017 or Stolk, 2019). While OUP has strong relations with some partners in ELEXIS, NUIG is also a partner of this project and Thierry Declerck (DFKI) is involved as an external expert in ELEXIS. 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 by the main technological partners of Prêt-LLOD, but also giving feedback on certain issues.

Cooperation with the ELEXIS infrastructure in the field of Digital Humanities resulted in the joint organisation of a workshop at the Digital Humanities conference in Utrecht (July 2019), on which we report in section 4.4 in this deliverable.

Cooperation with ELEXIS is also concerned with the discussion on standards, for example establishing bridges between OntoLex-Lemon,¹¹ as the result of a W3C Community Group,

¹⁰ See <https://lexdhai2019.acdh.oeaw.ac.at/> for more details.

¹¹ See <https://www.w3.org/2016/05/ontolex/> for the full specification of the model.



and the TEI Lex-0 encoding guidelines in development within the TEI community.¹² In this context a special meeting between the two communities was co-organized by the Prêt-à-LLOD partner NUIG as a satellite event to the eLex 2019 conference in Sintra, Portugal (see section 4.5 in this deliverable). At the same time, we could notice that the OntoLex-Lemon model was often mentioned in the papers submitted to eLex 2019.

Additionally, the Prêt-à-LLOD partner DFKI is in contact with the OASIS standardisation community, discussing the possible creation of a charter on Lexicographic Infrastructure Data Model and API.

6 Relations of Prêt-à-LLOD to the WordNet Community

The WordNet community is an important group of users and producers of language data, delivering so-called synsets that are a representation of lexical semantics properties associated with words. Prêt-à-LLOD is aiming at increasing the representation of WordNets data in the formats at the core of the Linguistic Linked Open Data cloud. John P. McCrae (NUIG) is a board member of the Global WordNet Association¹³ and both John McCrae and Thierry Declerck (DFKI) attended the Global WordNet Conference in July 2019 in Wroclaw, presenting different works on integrating WordNets in the OntoLex-Lemon model (McCrae et al., 2019b; Declerck et al., 2019b). NUIG is actively involved in the work of the WordNet community, firstly through the development of the Collaborative Interlingual Index¹⁴ and secondly through the development of the English WordNet¹⁵. DFKI started a new working group on multimodal WordNets¹⁶.

As such, Prêt-à-LLOD project aims have been disseminated through this community and the members of the community are active stakeholders in developing linguistic linked open data. Within the WP3 “Transforming, Linking and Workflows for Language Resources” of Prêt-à-LLOD active work is dedicated to the transformation of data included in the Open Multilingual WordNet initiative¹⁷, as this can be read in Deliverable 3.1 “Research Challenges Report v1” (section 2.2 Case study: Open Multilingual Wordnet in TSV).

7 Relations of Prêt-à-LLOD to the Terminology Community

Multilingual terminologies are an important component of the use cases defined for Prêt-à-LLOD. In order to include such terminologies in the Linked Data based workflows of

¹² TEI stands for Text Encoding Initiative (<https://tei-c.org/>). For TEI Lex-0, which is used more in the DARIAH community, see <https://dariah-eric.github.io/lexicalresources/pages/TEILex0/TEILex0.html>.

¹³ See <http://globalwordnet.org>.

¹⁴ <https://github.com/globalwordnet/ili>

¹⁵ <https://github.com/globalwordnet/english-wordnet>

¹⁶ See <http://globalwordnet.org/resources/working-groups/>

¹⁷ See <http://compling.hss.ntu.edu.sg/omw/> for more details.

the use cases, the project is also making use of the OntoLex-Lemon model. This model supports the use of the SKOS¹⁸ vocabulary for linking terms listed in terminological bases to a full lexical description to be encoded in the corresponding elements of OntoLex-Lemon. The cooperation with the terminology community is therefore an important aspect of the community building activities of the project.

One aspect concerns the presentation of Prêt-à-LLOD activities in conferences having a focus on terminology. This was the case for example for the TOTh 2019 conference¹⁹ where the Prêt-à-LLOD partner presented work on how to interlinking terminology and lexicography data. At TOTh 2019 we could also notice that OntoLex-Lemon is well received in the terminology community, as a certain number of presented works are referring to it, and this was also documented by post conference mail exchanges.

The relevance of the terminology community for Prêt-à-LLOD is also documented by cooperation established with European Institutions, as described in the next section.

8 Relations of Prêt-à-LLOD to European Initiatives

In the context of the Connecting Europe Facility (CEF) funding scheme, service contracts have been established for the collection, processing and storing of language resources that can support the quality of machine translation between European entities, as provided by eTranslation, whose aim is to “enable multilingual public services and communication” (<https://ec.europa.eu/cefdigital/wiki/display/CEFDIGITAL/eTranslation>). One such service contract oversaw setting up a portal, called ELRC-SHARE (<https://elrc-share.eu/>), in which relevant resources for eTranslation are described and stored, ideally using standard representations like TMX²⁰ (for aligned bilingual corpora) or TBX²¹ for (possibly multilingual) terminologies.

Prêt-à-LLOD can as a first step take advantage of terminological resources stored in ELRC-SHARE, and which are related to the use cases of the project. At the same time Prêt-à-LLOD will disseminate those rich language resources in the Linguistic Linked Open Data cloud (<http://linguistic-lod.org/lod-cloud>), in accordance with licensing conditions. As a second step, Prêt-à-LLOD will submit the language resources that the project has gained in its use cases to ELRC-SHARE, possibly supporting improvements of the eTranslation platform for specific domains.

In the context of defining possible cooperation with CEF activities, Paul Buitelaar (NUIG) presented Prêt-à-LLOD at the “Workshop on Showcasing Language Technology in H2020

¹⁸ SKOS stands for “Simple Knowledge Organization System” and is being used by a large number of terminologies, taxonomies and thesauri. See <https://www.w3.org/2004/02/skos/> for more details.

¹⁹ http://toth.condillac.org/wpcontent/uploads/2019/03/TOTh_2019_Program_27_03_2019.pdf

²⁰ TMX stands for “Translation Memory eXchange” and is an XML based format. “The purpose of the TMX format is to provide a standard method to describe translation memory data that is being exchanged among tools and/or translation vendors, while introducing little or no loss of critical data during the process.” (<https://www.gala-global.org/tmx-14b>).

²¹ TBX stands for is a XML-based standard for representing and exchanging information about terminology. See <https://www.tbxinfo.net/>



and CEF Telecom projects' in Brussels on April 26th, 2019. The workshop, organized by the Belgian translation automation consultancy 'CrossLang' in cooperation with DG Connect, brought together 17 ongoing H2020 and CEF Telecom projects.²² Related to this activity is the Innovation event organized by the Prêt-à-LLOD partner Semantic Web Company, where the CEO of CrossLang was invited to present the CEF funded market study for language technologies in Europe (see the section “Industry Events” in this deliverable).

We also think that Prêt-à-LLOD can directly cooperate with the European Open Data Portal (<https://data.europa.eu/euodp/en/home>), a possibility the project will investigate in its second year, when enough language and knowledge data have been collected and transformed into RDF-based vocabularies.

Representatives of Prêt-à-LLOD participated in a two days' workshop on "Semantic Interoperability for the Multilingual Web" organized by the Publications Office of the European Union (4th and 5th June 2019 in Luxembourg). This workshop was organised in the context of two actions in the domain of semantic interoperability that are funded by the ISA² programme of the European Commission (https://ec.europa.eu/isa2/home_en), and we quote here from the invitation letter:

- “Public Multilingual Knowledge Management Infrastructure for the digital single market (PMKI), aiming to create a platform for sharing and aligning of multilingual terminologies that are made available by different contributors to support public organisations, academics and businesses, in particular the language technology industry, with the implementation of the multilingual solutions that are needed to overcome language barriers within the European Union and to increase interoperability in general .”²³
- “VocBench, a multilingual platform for collaborative thesaurus management including the support of OWL ontologies, SKOS (XL) thesauri, generic RDF datasets, as well as Ontolex/LEMON resources .”²⁴

We are very happy to see here that the Publication Office are considering an extensive use of the OntoLex-Lemon model for the building of its multilingual Knowledge Bases and Prêt-à-LLOD is very keen to contribute to this effort, leading to an additional sustainability result of the LLOD community, and also helping the project crucially in their use cases, especially for the eGovernment use case.

9 Industry Events

As stated in the introductory section, Prêt-à-LLOD aims to present its methodologies and results at industry events. For this, the project is also committed to organize an “Innovation

²² Further details on the workshop and links to all presentations are available from <https://ec.europa.eu/cefdigital/wiki/display/ETCOMMUNITY/Workshop+Showcasing+LT+Agenda#WorkshopShowcasingLTAgenda-Presentations>.

²³ https://ec.europa.eu/isa2/actions/overcoming-language-barriers_en

²⁴ <https://joinup.ec.europa.eu/solution/vocbench3/about>

event” per year. In the following sections, we summarize this type of community building activities conducted by the project.

9.1 The first Prêt-à-LLOD Innovation Event: The ELSE-IF workshop @ SEMANTiCS 2019

The Prêt-à-LLOD partner Semantic Web Company organized the first Innovation event of the project as a satellite workshop to the Semantic 2019 conference, which was held in Karlsruhe, Germany, September 9. The Website of the ELSE-IF workshop: <http://www.pret-a-llod.eu/else-if19/>

The workshop attracted 15 participants in total. The affiliations of the participants were, SAP, Siemens, Evosoft, L3S, Trinidata, and for Prêt-à-LLOD: Semalytix, Semantic Web Company and DFKI. Two invited speakers gave statements about the current market and infrastructure situation of Language Technologies in Europe (and beyond): Luc Meertens (Crosslang) and Georg Rehm (DFKI)

The choice of the invited speakers reflected the interest of Prêt-à-LLOD to establish strong bridges to the Connecting Europe Facilities (CEF) infrastructure (Luc Meertens coordinated for the Commission a study that positions CEF Automated Translation, a building block of the Connecting Europe Facility, in the European market for language technologies (LT)) and to the European Language Grid (ELG) infrastructure (Georg Rehm is the coordinator of this project, which has also to deal with market aspects of LT).

The two invited talks - one with market focus and one research focus - complemented each other. First Luc Meertens gave an overview of the language technology (LT) market. The market appears segmented, small size but growing. The key recommendations stress a lack of venture capital and a necessity of stronger infrastructure. In the second invited talk Georg Rehm focused on the strategy of the European Language Grid project to address the identified challenges. Namely, a single infrastructure for single digital language market will be prepared in ELG project. Clearly, Prêt-à-LLOD and other similar projects benefit from such a market place and will use it to make the results visible, discoverable and sustainable.

The talks by Thierry Declerck, Matthias Orlikowski, Artem Revenko demonstrated different real-life use case for the usage of the Prêt-à-LLOD resources and methodologies: usage for the extension of eGovernment, SEM and SWC pilots from the Prêt-à-LLOD project.

In a concluding roundtable discussion, the participants made their suggestions to improve the event next year and to increase the impact. Among the various suggestions, we think of co-locating the next Innovation event with one of the following venues:

- Workshop at META-FORUM 2020²⁵, where a significant industry participation, around 30-40% can be expected. Many European projects are disseminated at the event, including an exhibition slot for demonstration purposes.

²⁵ This would be the second edition of META-FORUM to be organized by the ELG consortium. See <https://www.european-language-grid.eu/meta-forum-2019/> for the 2019 edition of META-FORUM.

- Workshop at LT Innovate²⁶, where an industrial focus and hence a more target audience can be expected.

It was acknowledged that for the next Innovation event it would make sense to focus on a tutorial in order to explain to the industrial participants how the language resource might be used to provide benefits for the industrial use cases. A cooperation with ELG is envisaged in this case.

9.2 Prêt-à-LLOD at the BDV PPP Meetup 2019

Representatives of the Prêt-à-LLOD partners UPM and DFKI have presented aspects of the project of the Linked Data infrastructure at the event organised by the Big Data Value Association (BDVA) in Riga, June 26-28 (<http://www.bdva.eu/node/1217>). One presentation was offered in the session “Recent advances in Artificial Intelligence for Natural Language Processing” and one in the session “Interoperability as an enabler of the data economy”. The company Tilde²⁷ and Prêt-à-LLOD aimed during those sessions at increasing the awareness of the relevance of language technologies in the context of Big Data applications.

During the event, a cooperation with the project “Special” (“Scalable Policy-aware Linked Data Architecture For Privacy, Transparency and Compliance“, <https://www.specialprivacy.eu/>) has been established, especially in the context of W3C activities dealing with privacy issues.

9.3 Prêt-à-LLOD and ELG jointly organized the “Initial Workshop on the Language Technology Market and Components Taxonomy” at the LT-Innovate Summit 2019

The Prêt-à-LLOD partner Semantic Web Company and the ELG partner Sail Labs jointly organized a workshop at the 8th Language Technology Industry Summit (24-26 June 2019, <https://www.pret-a-llod.eu/2019/07/03/involving-the-lt-industry/>). This event aimed at initiating a common effort on building an LT-related general usage taxonomy. Together with 45 representatives from European LT Industry, ideas for this usage taxonomy were discussed. This initiative is closely related to the one dealing with the further development of metadata for Language Technology resources, and herein lies a central cooperation between ELG and Prêt-à-LLOD for ensuring that industry is having an easy and clear access to relevant language technologies resources for their application.

10 The first Sustainability Event of Prêt-à-LLOD

Prêt-à-LLOD has defined a four levels strategy for ensuring sustainability: (1) through the form and representation of data, (2) through a fine-grained provenance, legal information,

²⁶ LT-Innovate is the Language Technology Industry Association (<http://www.lt-innovate.org/>) and organizes regularly a LT-Summit.

²⁷ <https://www.tilde.com/>

and smart policies for language data transactions, (3) through accessibility and repository management, and (4) at a community level.

The project started to talk with other projects in order to see which sustainability strategies are shared or can be considered as complementary, leading to an enlargement of the overall strategy for ensuring sustainability for language technologies. This first round of discussion was leading to the joint organisation by Prêt-à-LLOD and ELG of the session “Sustainability of the ICT-29 Projects” at META-FORUM 2019 dedicated to the sustainability approaches followed by the projects resulting from the call “ICT-29-2018: A multilingual Next Generation Internet”. There, John McCrae (NUIG) gave a talk jointly prepared by NUIG and DFKI on “Overview of the Sustainability Plans of the ICT-29b”, while Georg Rehm (coordinator of ELG) presented the sustainability plans of ELG leading to the creation of a legal entity. The session, which was attended by a large number of META-FORUM participants, has been very well received.

Parallel to the organisation of the META-FORUM session on sustainability, Prêt-à-LLOD contacted a number of infrastructure projects, like ELEXIS (in the field of eLexicography, see <https://elex.is/>), CLARIN-ERIC (in the broader field of Digital Language Resources to be made available to researchers in all fields, see <https://www.clarin.eu/>), DARIAH (Digital Research Infrastructure for the Arts and Humanities, see <https://www.dariah.eu/>), and for sure also ELG (<https://www.european-language-grid.eu/>). The motivation for this round of discussion is that Prêt-à-LLOD can also be considered as an infrastructure project, with a focus on Linked Data. And Linked Data are playing a role in all the larger infrastructure projects mentioned above, and there is a need therefore to adjust the distinct sustainability strategies to be developed for infrastructures in the field of LT. As a result of this round of discussion, Prêt-à-LLOD will investigate which aligned sustainability approaches should be evaluated and merged into one document. This will give the basis for the organisation of a second event dedicated to sustainability issues to be organised in 2020.

11 The Linked Data for Language Technologies (LD4LT) Community Group Report

The Linked Data for Language Technologies (LD4LT) is a community group that was founded in the previous FP7 project “LIDER” and that is used for the broader discussion of issues related to linked data and its applications in NLP. The group has fallen quite silent in the period after the end of the LIDER project and as a result we asked the former chairs of LD4LT if they would agree to have Prêt-à-LLOD members to be the new chairs of this group and direct it from a Prêt-à-LLOD project view. As such, we are now using this group to discuss some public activities, such as the META-SHARE OWL ontology²⁸, as well as disseminating project activities through its newsletters.

²⁸ <https://github.com/ld4lt/metashare/tree/dev>



12 Prêt-à-LLOD Standardisation Activities

The use of and the contribution to standards is a way to ensure sustainability of results (algorithms, data, formats, guidelines, etc.) delivered by a project. Prêt-à-LLOD is particularly involved in such standardisation activities, having as a focus the further development of modules for the core model OntoLex-Lemon, developed in the context of a W3C Community Group (<https://www.w3.org/2016/05/ontolex/>). For this version of the deliverable reporting on Task 5.4 activities, we just refer to the deliverable D5.1 “Report on Vocabularies for Interoperable Language Resources and Services”, in which the relevant existing standards are described in details, but also the activities that have been recently initiated by partners of Prêt-à-LLOD. As there will be no updates of D5.1, further standardisation activities by Prêt-à-LLOD will be described in the next iterations of D5.4 (D5.4.2 and D5.4.3).

13 Contributions to the new COST Action CA18209 - European network for Web-centred linguistic data science (“NexusLinguarum”)

We quote first from the Action’s web page (see <https://www.cost.eu/actions/CA18209>):
“The main aim of this Action is to promote 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 order to support the study of linguistic data science in the most efficient and productive way, the construction of a mature holistic **ecosystem of multilingual and semantically interoperable linguistic data** is required at **Web** scale. Such an ecosystem, unavailable today, is needed to foster the systematic cross-lingual discovery, exploration, exploitation, extension, curation and quality control of linguistic data. We argue that linked data (LD) technologies, in combination with natural language processing (NLP) techniques and multilingual language resources (LRs) (bilingual dictionaries, multilingual corpora, terminologies, etc.), have the potential to enable such an ecosystem that will allow for transparent information flow across linguistic data sources in multiple languages, by addressing the semantic interoperability problem.”

Therefore, the role of LD to support NLP will be core in this Action, as it is in Prêt-à-LLOD. In fact, there are two Working groups in the network that will explicitly study this: WG 1 - Linked data-based language resources and WG 2 - Linked data-aware NLP services. NexusLinguarum (as any other COST Action) will not receive funds for research but for networking activities. Through such networking activities (research visits, conferences/workshops organisation, scientific meetings, training schools, etc.), the Prêt-à-LLOD participants will be put in connection with a high number of other researchers and practitioners all across Europe. This will constitute an invaluable opportunity to further exchange ideas, discover new datasets of interest to the project (particularly for minority languages), as well as to discover application scenarios to the outcomes of Prêt-à-LLOD.



The joint organisation of several events between NexusLinguarum and Prêt-à-LLOD is also foreseen.

Currently, 35 COST countries have joined NexusLinguarum, and more will join the Action in the near future as Near Neighbour Countries or as International Partners (e.g., Georgia, Bielorrussia, USA). The Management Committee consists of more than 100 members and substitutes. The first working meeting of the NexusLinguarum will take place at the end of January 2020.

Jorge Gracia (UNIZAR) was the main proposer of NexusLinguarum and will coordinate the Action. John McCrae (NUIG) is the vice chair of NexusLinguarum, and Thierry Declerck (DFKI) was elected as its manager for scientific communications. With those persons occupying those positions in NexusLinguarum, we expect a high level of synergy between the types of activities foreseen in the COST Action and the research and development goals of Prêt-à-LLOD.

14 Conclusion

In this deliverable we reported on the community building strategy of Prêt-à-LLOD and listed the various actions taken by the project to extend the community of potential users and contributors to the Linguistic Linked Open Data infrastructure. We described the various communities with which we established or enhanced cooperations.

We also reported on the first Innovation and Sustainability events organized by Prêt-à-LLOD.

Together with the report D5.1 and the report on Dissemination activities (D6.1.1), we see the project on a very good way for reaching the goals specified in Work Packages 5 and 6.

15 References

Thierry Declerck, John P. McCrae. Proceedings of the Poster Session of the 2nd Conference on Language, Data and Knowledge (LDK-PS 2019), <http://ceur-ws.org/Vol-2402/>

Thierry Declerck, Melanie Siegel, Dagmar Gromann. OntoLex as a possible Bridge between WordNets and full lexical Descriptions. Proceedings of the 10th Global WordNet Conference, (2019b).

Thierry Declerck, Carole Tiberius, Eveline Wandl-Vogt Encoding lexicographic Data in lemon: Lessons learned_in: John P. McCrae, Francis Bond, Paul Buitelaar, Philipp Cimiano, Thierry Declerck, Jorge Gracia, Ilan Kernerman, Elena Montiel Ponsoda, Noam Ordan, Maciej Piasecki, Jan Wieczorek (eds.): Proceedings of the LDK workshops: OntoLex, TIAD and Challenges for Wordnets, Galway, Ireland, CEURS, 8/2017

Maria Eskevich, Gerard de Melo, Christian Fäth, John P. McCrae, Paul Buitelaar, Christian Chiarcos, Bettina Klimek, Milan Dojchinovski (Eds.). Proceedings of the 2nd Conference on



Language, Data and Knowledge (LDK 2019). OASlcs, Vol 70. Schloss Dagstuhl – Leibniz-Zentrum für Informatik GmbH, Dagstuhl Publishing, Saarbrücken/Wadern, Germany. Online available at <https://www.dagstuhl.de/dagpub/978-3-95977-105-4>.

Maria Gavrilidou, Penny Labropoulou, Elina Desipri, Stelios Piperidis, Haris Papageorgiou, Monica Monachini, Francesca Frontini, Thierry Declerck, Gil Francopoulo, Victoria Arranz, Valerie Mapelli. (2012) The META-SHARE Metadata Schema for the Description of Language Resources. Proceedings of the Eighth International Conference on Language Resources and Evaluation (LREC-2012).

Julia Bosque-Gil, Dorielle Lonke, Jorge Gracia, Ilan Kerneremann. Validating the OntoLex-lemon Lexicography Module with K Dictionaries' Multilingual Data. Proceedings of Sixth Biennial Conference on Electronic Lexicography, eLex 2019, (2019).

John P. McCrae, Alexandre Rademaker, Francis Bond, Ewa Rudnicka and Christiane Fellbaum English WordNet 2019 — An Open-Source WordNet for English. Proceedings of the 10th Global WordNet Conference, (2019b).

John P. McCrae, Carole Tiberius, Anas Fahad Khan, Ilan Kernerman, Thierry Declerck, Simon Krek, Monica Monachini and Sina Ahmadi. The ELEXIS Interface for Interoperable Lexical Resources. Proceedings of Sixth Biennial Conference on Electronic Lexicography, eLex 2019, (2019).

Sander Stolk. A Thesaurus of Old English as Linguistic Linked Data: Using OntoLex, SKOS and lemon tree to Bring Topical Thesauri to the Semantic Web. Proceedings of Sixth Biennial Conference on Electronic Lexicography, eLex 2019, (2019).

