Prêt-à-LLOD

D4.1
Business Pilot Specification

Author(s):

Christian Blaschke, Maria Khvalchik, Artem Revenko, Guilherme Rodrigues, Roser Saurí, Meritxell González, Khalil Ahmed, Eva Theodoridou, Deirdre Lee, Katharine Cooney, Mario Romera, Matthias Orlikowski, Susana Veríssimo, Soufian Jebbara, Maria Pia di Buono, John McCrae, Matthias Hartung

Date: June 28, 2019

This project received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 825182. The information and views set out in this publication are those of the author(s) and do not necessarily reflect the official opinion of the European Union.

H2020-ICT-29b
Grant Agreement No. 825182
Prêt-à-LLOD - Ready-to-use Multilingual Linked Language Data for Knowledge Services across Sectors

D4.1
Business Pilot Specification

Deliverable Number: D4.1
Dissemination Level: Public
Delivery Date: June 30, 2019
Version: 1.0
Author(s): Christian Blaschke, Maria Khvalchik, Artem Revenko, Guilherme Rodrigues, Roser Saurí, Meritxell González, Khalil Ahmed, Eva Theodoridou, Deirdre Lee, Katharine Cooney, Mario Romera, Matthias Orlikowski, Susana Veríssimo, Soufian Jebbara, Maria Pia di Buono, John McCrae, Matthias Hartung

Document History

Version Date	Changes	Authors
0.1	**First draft**	Christian Blaschke, Maria Khvalchik, Artem Revenko, Guilherme Rodrigues, Roser Saurí, Meritxell González, Khalil Ahmed, Eva Theodoridou, Deirdre Lee, Katharine Cooney, Mario Romera, Matthias Orlikowski, Susana Veríssimo, Soufian Jebbara, Matthias Hartung
0.2	**Revisions after review**	John McCrae, Maria Pia di Buono

1.0	**Final Version**	Christian Blaschke, Maria Khvalchik, Artem Revenko, Guilherme Rodrigues, Roser Saurí, Meritxell González, Khalil Ahmed, Eva Theodoridou, Deirdre Lee, Katharine Cooney, Mario Romera, Matthias Orlikowski, Susana Veríssimo, Soufian Jebbara, John McCrae, Maria Pia di Buono, Matthias Hartung

Table of Contents

1 Preamble: Background and Motivation

Prêt-à-LLOD aims to create data value chains that can be used across industrial sectors in order to reduce development and time to market for multilingual language-technology-based software products and applications, while increasing interoperability between multilingual datasets and multilingual language technology services. This involves the challenges of discovery, transformation, linking and composition of language resources, for which the project will provide specific workflows and methodologies.

Their transferability and applicability to practical language technology solutions and applications will be demonstrated in four industry pilots to be carried out by industry partners as summarized in Table 1:

Pilot ID	Pilot Description	Industry Partner in Charge	Reference Section in this Document
I	Multilingual Knowledge Graphs for Knowledge Management across Sectors	Semantic Web Company (SWC)	Section 2
II	Linking Lexical Knowledge to Facilitate Rapid Integration and Wider Application of Lexicographic Resources for Technology Companies	Oxford University Press (OUP)	Section 3
III	Supporting the Development of Public Services in Open Government both within and across borders	Derilinx (DLX)	Section 4
IV	Multilingual Text Analytics for Extracting Real-World Evidence in the Pharma Sector	Semalytix (SEM)	Section 5

Table 1: Overview of pilot projects described in this document

While Prêt-à-LLOD workflows and methodologies cut across many potential application domains and sectors, pilots will showcase potentials in the context of the following sectors specifically:

Technology Companies: As large technology companies expand further beyond their traditional markets, their demand for multilingual dictionary content and lexical materials increases in parallel. These companies would like to extend the services already available in English, French, Spanish to the other EU working languages and beyond, but the content to support these is not readily available. At **OUP** this gap is currently addressed by creating and gathering lexical knowledge in many languages, from many sources. Prêt-à-LLOD will facilitate this market-leading dictionary data to become much more interlinked and manageable.

Pharma: In times of value-based contracts and mixed pricing models, real-world evidence has been identified as a key success factor for pharma companies. Real-world evidence is evidence for the effectiveness and safety of a drug product, gathered outside of the controlled settings of a clinical trial, in order to provide a proof of added value of a drug in specific patient populations. Extracting real-world evidence requires to analyze large volumes of heterogeneous data, which also includes the subjective assessments of patients and doctors on the added value of a drug. Such content is typically available as unstructured natural language text in multiple languages and having access to better solutions for extracting evidence from multilingual content would have an important impact and a strong market potential. Prêt-à-LLOD workflows and methodologies will be adopted by **Semalytix** in order to develop multilingual text analytics services and applications that support generating real-world evidence by automatically analysing multilingual data from patient forums, social media, CMS data, or other textual sources.

Government Services: Due to linguistic fragmentation, many citizens and businesses cannot directly engage with online services provided by public administrations. Hence, there is a clear need for advancements in human language technology in the public sector. In Prêt-à-LLOD, we address the dual challenges of (i) providing cross-border public services, essential to achieve an inclusive Digital Single Market, and (ii) the portability of public services and knowledge sharing across jurisdictions for improved collaboration and cost savings. Furthermore, many European countries and cities are collaborating to achieve the Digital Single Market, including innovative solutions for sectors such as energy, transport, housing and infrastructure. To demonstrate the strong need for language technologies to facilitate the availability and accessibility of cross-border data and public services, **Derilinx** will deploy the Prêt-à-LLOD technology stack to provide enhanced functionality in the interpretation of user queries to government health services and improved access to cross-border Open Data.

In addition, **Finance** will be in scope of the pilot carried out by **Semantic Web Company** (which is conceived as genuinely cross-domain), and an optional secondary domain in the **Semalytix** pilot.

As overarching challenges, all pilots will be addressing facets of *cross-language transfer* or *domain adaptation*, albeit in varying degrees.

In Pilot I, Semantic Web Company aims at improving term extraction and concept matching services as offered by their flagship product, PoolParty. While PoolParty is designed to be applicable to multiple domains, the goal is to extend term extraction and concept matching workflows currently existing in PoolParty. In three sub-pilots, particular attention will be paid to aspects of quality enhancement in term extraction, improved lemmatization capabilities in concept matching, improved disambiguation capabilities as a prerequisite for concept matching, and extension to several new languages.

Oxford University Press, as a provider of highly sophisticated, comprehensive and lexically rich resources for the language technology industry will devote their activities in Pilot II to linking lexical data for language services, thus directly addressing one of the key challenges in focus of 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 sense level, and linking corpus data to dictionary senses by means of word sense disambiguation. In the latter aspect, Pilots I and II share mutual goals, while differing in terms of 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 as for which no dictionaries with sense-level information are currently existing). As another commonality, the work in Pilot II will also enable to port existing lexical resources to new languages.

In Pilot III, Derilinx aims to provide tools and interfaces for intuitive and cross-border access to open data portals using natural language. In the first of two sub-pilots, a chatbot will be developed; this involves mapping natural-language user queries into formal queries against an open data health service portal. In order to achieve this, Derilinx will take a previously developed chatbot for an open government portal and transfer it to the health domain. In a second sub-pilot, cross-language transfer techniques will be applied in order to provide a web interface for users to access cross-border open data portals in their native language.

In Pilot IV, Semalytix 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 customers from the pharmaceutical industry. Similarly to Pilot III, the high degree of domain-specificity of existing mono-lingual analytics components implies that cross-lingual transfer approaches based on available LLOD resources and services will likely require efforts for domain adaptation. If time and resources permit, the developed framework for domain-preserving cross-lingual transfer will be subsequently employed to introduce finance as an additional domain into the Semalytix analytics stack.

As outlined in the Statement of Work, business pilots will be implemented subsequent to the planning phase (M1-6) throughout M6-36 of the project. Deliverables per each pilot project are two Pilot Reports (first version at M24, final version at M36, respectively).

The objective of the present document is to provide a comprehensive specification of the pilots. This involves, for each individual pilot project, the following aspects:

- Definition of use cases and user stories as elicited by the industry partners from their clients and business contacts
- Specification of functional and non-functional requirements
- Definition of milestones during pilot execution
- Development of appropriate benchmarks for validation and testing
- Evaluation plan

These aspects, for each pilot, will be detailed in the sections below, reflecting the current state of planning and agile implementation at M6 of the project.

2 Pilot I: Multilingual Knowledge Graphs for Knowledge Management Across Sectors (SWC)

PoolParty® Semantic Suite[1], SWC's flagship product, is an AI platform based on semantic technologies and machine learning. It helps organizations to build and manage knowledge graphs as a basis for various AI applications. As a semantic middleware, PoolParty extracts the semantic meaning from data and links business objects and content assets automatically.

PoolParty Extractor (a part of the PoolParty Semantic Suite) supports highly scalable and precise entity extraction, based on knowledge graphs as well as machine learning, which can be combined, put in series, or even used as parts of more complex rules and constraints for sophisticated text mining tasks. Its ability to transform structured and unstructured information into RDF offers new options for data analytics.

The goal of the pilot in Prêt-à-LLOD is to improve some concrete aspects of term extraction (can be understood as named entity recognition) and concept matching (can be understood as tagging text with concepts from an existing vocabulary).

1. Use of linguistic methods for improving the quality of extracted terms (pilot Ia).
2. Improvement of the lemmatization capabilities for improving concept matching (pilot Ib).
3. Improvement of the concept disambiguation capabilities (pilot Ic).

Pilot Ia, improving the quality of extracted terms: the goal is to introduce a more sound linguistic approach (like part-of-speech tagging, phrase chunking and others) for the extraction of terms with the expectation that this will increase the quality of the terms. This has an impact in different parts of the product. On the one hand this affects extraction results on processed documents, as better terms lead to better annotation results. And on the other hand in the corpus analysis feature of PoolParty where better term extraction makes users more effective in finding suggestions for new concepts that should be added to a domain thesaurus.

Pilot Ib, improving domain specific concept extraction: the goal of this is to extend the lemmatization capabilities of PoolParty and make them adaptable to new domains. The expectation here is to decrease the level of missed concept annotations in the cases where domain specific word forms appear in a text that are not covered in the tagging vocabulary.

Pilot Ic, improving concept disambiguation: here the goal is to improve the current implementation of concept disambiguation in PoolParty. What we want PoolParty to be able to do is to recognise when terms or concepts appear in different senses in documents, and

[1] https://www.poolparty.biz/, https://help.poolparty.biz/pp7/white-papers-release-notes/poolparty-technical-white-paper

to train the system on instances of known senses so that it can distinguish them in the annotation process.

2.1 Pilot Ia: Improve the quality of extracted terms

Term extraction in PoolParty is currently based on (stop) word lists, pruning of generated terms and distribution statistics over text corpora. The goal is to use linguistic clues to improve the quality of terms and scoring in areas such as concatenation of phrases (e.g. verb and noun phrases) and suboptimal segmentation of multiword terms.

2.1.1 Objectives

The goal of this development is to increase the quality of terms that are extracted from different components in PoolParty. Success can be measured in two ways. On the one hand side on a pure data level, i.e. extract terms with different methods and assess by some criteria if the quality is different. And on the other hand there is the effect that those results have on the actual user of PoolParty, i.e. if and how terms that are produced by different methods have an influence on how effectively a user can work with PoolParty.

2.1.2 Use Cases/User Stories

User stories for features where term extraction is involved:

- Exclude verbs in terms that are not nominalised (in which case they would form part of a noun phrase)
- Reduce level of fragmentation of terms - tiger vs. tiger shark - vs. tiger shark fishing

2.1.3 Data Sources

PoolParty covers a range of languages and the goal is to improve the workflow as many languages as following set: English, German, Spanish, French, Czech, Slovak, Dutch and Russian. The final set depends on the capabilities of the used resources.

The tools of interest are Part-of-Speech (PoS) taggers which assign parts of speech to each word as well as Chunking tools capable of retrieving multiple word phrases. We have identified following tools based on different criteria such as easiness of usage, language coverage, licencing, usage in production, API availability etc.

PoS taggers and chunking tools:

1. **Stanford CoreNLP** is a widely used Java-based NLP toolkit from Stanford University with GNU General Public License. Covers English, German and Spanish.
2. **Apache OpenNLP** is a Java-based NLP toolkit made by Apache Software Foundation and licenced with Apache Licence. It covers English, German, Spanish and Dutch.

3. **The Natural Language Toolkit (NLTK)** is a Python-based suite of libraries with Apache Licence. Covers English, German, Spanish, French, Dutch and Russian.
4. **Spacy** is a relatively new Python-based open-source NLP library with an MIT licence. It covers English, French, Spanish and Dutch.

2.1.4 Requirements

Functional Requirements

The objective is to develop a method that chunks phrases, specially noun and verb phrases, in a text document. Two areas in PoolParty have to be taken into account.

PoolParty Extractor:

- Extract terms based on complete phrases
- Add phrase type to extracted terms and allow filtering of results by type
- Filter concept annotations (based on vocabularies) and remove annotations that do not overlap with at least one noun phrase

PoolParty corpus analysis:

- Apply the same filtering of concept annotations as for the PoolParty Extractor
- Term extraction should be based on phrases and separated by phrase PoS
- Develop a method for noun phrases that detects true nested-ness of terms that avoids splitting named entities. E.g. in a text with "tiger shark" the term "tiger" is not a true named entity (for that text), but in "tiger shark fishing" the terms "tiger shark" and "fishing" are valid terms. This should be detected based on phrase distributions in the corpus and the association of phrase heads with different terms.
- Adjust term scoring by linguistic criteria such as the likelihood that a term corresponds to a true named entity in the corpus

Non-Functional Requirements

- Creating phrases of the text of a typical A4 page (this corresponds usually to about 500, see https://anycount.com/WordCountBlog/how-many-words-in-one-page/) may not take more than 500ms. Target value should be below 100ms.
- The programming language of the solution should be Java or an easy integration into Java should be available.
- The available license agreements should either allow integration into PoolParty without requiring PoolParty code to be made open source (e.g. Apache or MIT license) or a commercial license needs to be available.

2.1.5 Milestones

Milestone ID	Goal	Due date (Project Month)
MS-I.1a	Integration of PoS taggers into PP workflows	M12
MS-I.2a	Implementation/integration of chunking mechanism	M17
MS-I.3a	Upscoring or filtering of NP and nouns in the output	M22
MS-I.4a	Quantitative Experiment	M30
MS-I.5a	Qualitative Experiment	M34

2.1.6 Deliverables

Report of the quantitative and qualitative experiment

Contribution to KPIs 4.1, 4.4

2.1.7 Benchmarks and Evaluation

The goal of this development is to increase the quality of terms that are extracted in different components in PoolParty. Success can be measured in two ways. On the one hand side on a pure data level, i.e. extract terms with different methods and assess by some criteria if the quality is different. And on the other hand there is the effect that those results have on the actual user of PoolParty, i.e. if and how terms that are produced by different methods have an influence on how effectively a user can work with PoolParty.

The first approach is based on corpus analysis of larger sets of documents. As quality of extracted terms is difficult to measure we will make an indirect assessment that measures **term rankings of different methods**. A corpus with a valid reference vocabulary is needed for that, i.e. a vocabulary that has a good coverage of the technical terminology used in the corpus. How the evaluation works is that one extracts terms from the corpus and sorts those terms according to their score (over the corpus). Then batches are taken of equal number of terms starting at the top of the list. E.g. 10 batches of 100 terms each, terms 1 - 100 in the first batch, terms 101 - 200 in the second batch and so on. Terms in each batch are looked up in the reference vocabulary to get the number of terms from each batch that appear there. What one should see is that the number of "valid" terms decreases the further one moves down the list. The evaluation then consists of running this experiment with different term extraction methods and compare the results. Better methods should generate more terms that match the reference vocabulary at the top of the list than lower performing methods.

Secondly, we plan to design **user acceptance tests** and measure the effect that terms that are generated by different methods have on the user experience. Extracted terms appear in two places in PoolParty, the corpus analysis and the annotation of individual documents with the PoolParty Extractor. The first step would be to show users different results and ask them to express their preferences towards the different options. This is a purely qualitative measure. For document level extraction it is difficult to design a more quantitative evaluation, but for corpus analysis our intention is to go deeper and create tasks to allow us to measure time to completion towards a goal. What corpus analysis does is to provide terms from a corpus for a user to create or enrich a thesaurus with them. The task we intend to set-up is to let users create thesauri from lists of extracted terms that were generated with different methods and see if there are significant differences in the time it takes to create a thesaurus to a certain level of completion.

The type of data we are going to use is written data such as laws, descriptions and news. Currently we are not interested in social media or speech data sources. Both domain specific and general domain will be presented.

Resources for benchmarking PoS taggers: https://universaldependencies.org/
Resources for benchmarking chunkers: https://github.com/boudinfl/ake-datasets

2.2. Pilot Ib: Domain specific lemmatization

Lemmatization is used in PoolParty to normalise terms and in the detection of concepts (from a thesaurus). Only a limited range of languages is currently covered and the goal is to extend that. The intention is to implement corpus learning of lemmas so users of PoolParty can improve lemmatization for their domain with the ultimate goal of improving the coverage of concept matching (i.e., annotation of concepts from a vocabulary to text) by better bridging the gap of the surface forms contained in the vocabulary and what appears in the documents that need to be tagged.

2.2.1 Objectives

The lemmatization method in PoolParty is based on a fixed dictionary for each language that contains the corresponding pairs of base and inflected forms. With the development of domain specific lemmatization the PoolParty Extractor will be able to generate lemmas for unseen words.

2.2.2 Use Cases/User Stories

User stories for domain specific lemmatization:

- Semantic Web Company uses lemmatization and large corpus for a specific language and provides a base lemmatization model for a certain language.
- User can extend lemmatization model with a domain specific corpus that is added to the base model for a language.

2.2.3 Data Sources

Lemmatizer tools is the same set of tools as specified in Pilot 1a:

1. Stanford CoreNLP
2. Apache OpenNLP
3. Wordnet Lemmatizer with NLTK
4. Spacy

2.2.4 Requirements

Functional Requirements

- The following languages should be covered: English, German (with existing models to be improved), Spanish, French, Czech, Slovak, Dutch and Russian (new models to be created).
- 90% of unique words and 95% of unique verbs, adjectives, adverbs in a corpus should be covered (test sets to be defined).
- Performance should be reasonable also in technical domains like finance, engineering, biomedicine, … (i.e. the above specified numbers should be met).

Non-Functional Requirements

- Lemmatising all words in a typical A4 page (500 words) may not take more than 500ms. Target value should be below 100ms.
- The programming language of the solution should be Java or easy integration into Java should be available.
- The available license agreements should either allow integration into PoolParty without requiring PoolParty code to be made open source (e.g. Apache or MIT license) or a commercial license needs to be available.

2.2.5 Milestones

Milestone ID	**Goal**	**Due date (Project Month)**

MS-I.1b	Integration of lemmatizers into PP workflows	M18
MS-I.2b	Experiment	M30

2.2.6 Deliverables

Report of the experiment

Contribution to KPI 4.3

2.2.7 Benchmarks and Evaluation

The lemmatisation method in PoolParty is based on a fixed dictionary for each language that contains the corresponding pairs of base and inflected forms. With the development of domain specific lemmatisation the PoolParty Extractor will be able to generate lemmas for unseen words. The improvement that this brings will be evaluated in two ways. First again at the level of the data where the **number of lemmatised words** is determined by using the old vs. the new method. In a second step we plan to evaluate the **impact the lemmatisation has on concept detection**. I.e. we would try to find out if the new lemmatisation method leads to an increase in the number of concept detections for a given thesaurus and a document corpus, and if the additionally detected concepts are correctly annotated or false positives are introduced.

Resources for benchmarking http://www.meta-share.org/:

- English lexicon http://metashare.ilsp.gr:8080/repository/browse/british-english-source-lexicon-besl-version-22/dc410e62de6811e2b1e400259011f6eaff8112b159c346f8a910378af93ece2a/
- English lexicon http://metashare.ilsp.gr:8080/repository/browse/english-lexicon-with-morphological-information/97686fc0de7111e2b1e400259011f6eabb30be28f6ee4796a439f4f023fbfe72/
- English dictionary http://metashare.ilsp.gr:8080/repository/browse/new-oxford-dictionary-of-english-2nd-edition/9460637ede6b11e2b1e400259011f6ea58609ecf25e1458f8e72077ed6ad7a70/

2.3 Pilot Ic: Word sense induction and word sense disambiguation

Ambiguity is a phenomenon in natural language where words or terms can have different meanings (or senses). Our interest here is in terminologies rather than lexical aspects like senses of words. The PoolParty annotation system needs to distinguish different senses when it finds ambiguous terms in text. The problem that needs to be solved here consists of two parts, first to induce senses from a corpus, and second to train an annotator to distinguish those senses. Word sense induction from text corpus allows users to become aware of different meanings in their vocabulary and supports them in collecting training examples for individual senses for training a disambiguation model. This model is then used in the annotation process for deciding which sense is present.

2.3.1 Objectives

Word sense induction and disambiguation are new functionalities for PoolParty and they will allow new workflows that did not exist before in this way. We therefore plan to perform again an evaluation close to the data to validate the methods as such, and a qualitative assessment on how effective the new functionalities in PoolParty are for the users.

The following new functionalities will be realised based on the implementation of these methods:

1. Run a corpus analysis and the system shows which concepts potentially have multiple senses. The user can inspect the suggestions and split concepts if needed to reflect each sense.
2. The corpus analysis also extracts a list of terms where the existence of multiple senses can be indicated. Users can then create concepts for each sense.
3. Train the extractor to distinguish the different senses of concepts and annotating them correctly in text.

2.3.2 Use Cases / User Stories

User stories for word sense induction (WSI) and word sense disambiguation (WSD):

1. Train disambiguation model on pre-classified document sets
2. Incrementally improve trained disambiguation model
3. Induce senses in detected terms
4. Induce senses in concepts in thesaurus

2.3.3 Data Sources

1. RDF data based on the OntoLex-Lemon Model of English WordNet 2019
2. Wikilinks Dataset http://www.iesl.cs.umass.edu/data/data-wiki-links

2.3.4 Requirements

Functional Requirements

Develop a method for word sense induction that induces senses of terms from text corpus:

- Input is a set of documents
- First step is to extract (single and multi-token) terms
- Then method should detect for each term in the corpus if it appears in clearly different meanings
- Meanings are expressed in terms of their environment, i.e. the terms that occur around them in the text
- To each term the corresponding meaning is attached

Develop a method for word sense disambiguation that can be trained to distinguish meanings of terms in text:

- Input is a term and a set of documents where the term appears in different meanings
- The documents are annotated in the sense that for each document the meaning in which the term occurs is specified
- The method should produce a trained model that returns for an input document + term to be disambiguated the correct meaning

Non-Functional Requirements

The following time constraints should be met:

- Sense induction for a set of 100 documents of the length of a typical A4 page (500 words) where a term occurs in 5 different senses should be below 2 sec (measured for each term).
- Training a disambiguation model for a set of 100 documents of the length of a typical A4 page (500 words) where a term occurs in 5 different senses should be below 5 min (measured for each term).
- The disambiguation of one term in a document should not take longer than 100 ms. Better towards 20 ms.

The number of training instances for a method to produce meaningful results is important to make it practically viable, so we need to establish realistic numbers of training instances with which the methods still work "good enough" (to be established what that means exactly):

- Sense induction should work reasonably well with 10 documents per sense.
- Training a disambiguation model should work reasonably well with 20 documents per sense.

Prediction accuracy: it is difficult to establish overall accuracy levels that the methods should reach because performance will depend a lot on the situation like the amount of training data, the quality and consistency of the data, and there will be cases that are easier and others that are more difficult to work on. But the expectation is that for the typical case the F-score should be around 0.85 so that it makes sense to use these methods in practise.

2.3.5 Milestones

Milestone ID	**Goal**	**Due date (Project Month)**
MS-I.1c	Word Sense Induction	M17
MS-I.2c	Word Sense Disambiguation	M22
MS-I.3c	Quantitative Experiment	M27
MS-I.4c	Qualitative Experiment	M34

2.3.6 Deliverables

Report on experiments

Contribution to KPI 4.4

Industry-related publication

2.3.7 Benchmarks and Evaluation

Word sense induction and disambiguation are new functionalities for PoolParty and they will allow new workflows that did not exist before in this way. We therefore plan to perform again an evaluation close to the data to validate the methods as such, and a qualitative assessment on how effective the new functionalities in PoolParty are for the users.

The evaluation of the methods will be performed on a Wikipedia corpus and a parallel corpus of web pages that link to the Wikipedia pages (the wiki-links corpus). The links serve as a gold standard of what the correct annotation should be. During evaluation the corpus of web pages is used the links to the Wikipedia pages are hidden to the methods. The word sense induction method then needs to produce the correct groups of senses. The word sense disambiguation method is trained on (a subset of) the known senses and needs to predict the correct sense for unseen instances.

The following new functionalities will be realised based on the implementation of these methods:

- Run a corpus analysis and the system shows which concepts potentially have multiple senses. The user can inspect the suggestions and split concepts if needed to reflect each sense.
- The corpus analysis also extracts a list of terms where the existence of multiple senses can indicated. Users can then create concepts for each sense.
- Train the extractor to distinguish the different senses of concepts and annotating them correctly in text.

The evaluation will consist of asking users for their satisfaction with the new features.

3 Pilot II: Linking lexical knowledge to facilitate rapid integration and wider application of lexicographic resources for technology companies (OUP)

3.1 Overview

The Dictionaries division at OUP holds and maintains a set of highly curated dictionaries, rich in lexicographic content and detailed sense information, as well as corpus data used for a number of activities, including linguistic research in order to study language usage and trends, enhancing or updating dictionary content, or supporting the development of language models that feed into language processing technologies. Both, dictionaries and corpus datasets are annotated with several layers of information such as metadata on the type of the text (e.g., domain, publication date, genre, provenance, region, register) and linguistic features (e.g., lemmas, part of speech, phrases, translations, example sentences, etc.).

However, corpus and dictionaries remain unconnected among them, which dismisses a two-way opportunity of, on the one hand, enhancing dictionary content with additional data from corpora (e.g., adding new example sentences from real corpus text) while, on the other, extending corpus data with further annotation layers from dictionary hand-curated information (e.g., grammatical information, sense, etc.) that can then be used in language technology applications.

Linking dictionaries and corpus datasets at the level of meaning is a crucial component in order to be able to derive new content, both within the same language or across multiple languages, or to enrich currently available data.

Pilot II focuses on devising and developing technology for linking lexical knowledge in order to facilitate rapid integration of lexicographic resources and allow for wider application of these types of data for companies in the language technology area, such as multilingual search, cross-lingual document retrieval, domain adaptation, and lexical translation. In particular, the goal of this pilot is to explore and use state-of-the-art methods and techniques in computational semantics, data mining and machine learning for linking language data at the level of meaning. This pilot will free up and also connect data that often exists in disparate silos, thus leading to a significant overhead in managing, enriching and reusing of this data. In this respect, the work carried out here will directly contribute to Challenge 4, concerning the linking of conceptual and lexical data for language services.

Two tasks (or sub-pilots) are involved:

Pilot II.A Linking different dictionaries at the level of meaning; that is, at the level of sense, which in monolingual dictionaries is featured by definitions and in bilingual ones, by translations.

Pilot II.B Linking corpora to dictionaries at the level of meaning. This task involves tagging corpus data with dictionary senses, thus linking corpus text to the dictionary content. It is a task that falls within the area of word sense disambiguation.

The following subsections present objectives, KPI contributions and deliverables, which are shared by both sub-pilots. Then, we move to independent sections for each sub-pilot, detailing their specifics: use case enablers, data sources to be used, functional and non-functional requirements, planned milestones, and finally benchmarks and evaluation approach.

3.2 Objectives

Both sub-pilots revolve around the following goals:

1. **Having multilingual lexical content interlinked already available** to be able to immediately respond to (ad hoc) user requests that otherwise would require a long production and curation process (e.g., multilingual wordlists).
2. **Facilitating the creation of new lexical content from current datasets**. For example, developing lexical translations between languages pairs for which there is no bilingual dictionary available, by means of an intermediate dictionary acting as a hub to which other bilinguals covering the targeted languages are linked.
3. **Supporting the enhancement of current dictionary datasets.** Having dictionary and corpus content interlinked will facilitate dictionary enhancement of different sorts:
    a. **Through semi-automatic processes.** For instance, extending an information-scarce dictionary with data from a richer dictionary, or from corpus data (e.g., adding new example sentences), identifying inconsistencies between dictionary pairs, discovering differences in sense coverage between two dictionaries (i.e., missing senses), etc. We aim to set up a methodology that helps to achieve this purpose by automatic means while ensuring a high degree of quality in the results.
    b. **Through manual activity.** Corpus content linked to dictionaries at the sense level will contribute highly relevant evidence information for editorial curation, such as statistics on word trends from sense-tagged corpora (from sub-pilot II.B), which can help identify lexical entries to which editors should put some attention; similarly, quality estimates on dictionary sense mappings (obtained in sub-pilot II.A) can be used as an indicator of either inconsistencies in the data or low quality content, and therefore point to areas where manual work is required to achieve the high quality standards expected by OUP's customers.
4. **Supporting the enrichment of corpora with additional layers of annotations** (e.g., grammatical information hand-encoded in dictionaries, inflectional properties from a morphology dataset, sense tags, etc.), which can be used for developing different kinds of applications within the area of human language technologies.
5. **Improving efficiency, reducing costs.** Ultimately, we aim to devise automated ways that support the development of high quality lexicographic data in order to increase process efficiency and reduce production costs.

3.3 KPI Contributions

Nr	KPI		Benchmarks
4.1	Increase efficiency/speed-up adaptation to new domains and languages (development time / time to market)	30-50%	Compare to previous internally developed OUP products. Implementation time, cost, resources varies across initiative: - Time: year scale - Budget: up to £200000 - Resources: suppliers, freelancers (up to 65)
4.2	Average reduction in the time (person hours) to convert a resource to linked data at end of project.	90%	The development of a conversion pipeline should decrease the number of tasks and invested (human) time required to complete the development of each single source.
4.3	Increase of stakeholder cost savings/sales as a result of Prêt-à-LLOD technologies and solutions due to usage of free language resources (compared to projections following the average market trends)	Savings of 5,000 Euros per language per user	1. Benchmark estimation: - time/cost to find and gather data - time/cost to find NLP tools 2. Datasets benchmark 3. Product development
4.4	Satisfaction level of early adopters of pilots in a survey	4-5/5	Lexicographers satisfaction survey (internal users)when using results from the pilot

3.4 Deliverables

Work carried out in PIlot II will directly contribute to Challenge 4, concerning the linking of conceptual and lexical data for language services.
Deliverables for this pilot will consist of the following content:

- **Final report**, briefing on the activity carried out in each of the sub-pilots and reporting results.
- **Technical documentation** generated as part of the activity around the pilot: annotation guidelines, data formats, algorithms and software libraries used, architectural decisions, etc.
- **Contribution to KPIs** (cf. Section 3.3 above)
- **Datasets:**
  - Data samples to be linked in each sub-pilot (dictionary content, corpus text) together with the resulting links.
  - Data schemas supporting the data samples.
- **Publications** in scientific forums resulting from the activity in the pilot.

3.5 Pilot II.A: Linking dictionaries at the level of meaning

3.5.1 Use Cases Enablers

The ultimate goal of devising automatic means for eliciting connections between dictionaries and corpora at the sense level is to facilitate the creation of new content and the enhancement of current one. The latter tasks, however, fall out from the scope of the current project. Because of that, we consider that the drivers for this pilot should be considered more as enablers than use cases. The following are examples of such enablers:

Type	Use case	Example
Internal	Enrich current lexical content by complementary information from other dictionaries	Given a pair of linked dictionaries A-B, content from dictionary A can be used to extend, check inconsistencies and enhance dictionary B, and vice versa
Internal	Develop new bilingual dictionary content	If we have two bidirectional dictionaries (e.g., Chinese ↔ English and Korean ↔ English), we can generate a new bidirectional dictionary (Chinese ↔ Korean)
External	Support lexical translation	We are able to feed lexical translation systems with content on new language pairs.
External	Support multilingual search engines	We are able to feed multilingual lexical alignments for cross-lingual search
External	Support metadata tagging tools	We provide multilingual lexical alignments for tagging data (e.g., images) in different languages

3.5.2 Data Sources

1. **Dictionary datasets**
    - *Oxford Dictionary of English* (ODE)[2]
    - Oxford bilingual dictionaries[3]
        - English-German (EN-DE)
        - English-Spanish (EN-ES)
        - English-French (EN-FR)
        - English-Italian (EN-IT)
        - English-Russian (EN-RU)
        - English-Chinese (EN-ZH)
2. **Sense-link annotations**
    - Manual annotations between ODE and the English side of the bilingual dictionaries listed above.

[2] https://en.oxforddictionaries.com/ (August 2017 release)

[3] https://premium.oxforddictionaries.com/

    - Features used by the sense-linking system that generated those links, as well as confidence scores as obtained from it.
3. **Data schemas** (OUP proprietary)
    - *OxMono* (for monolingual dictionary content)
    - *OxBiling* (for bilingual dictionary content)
    - *LeXML* (for monolingual and bilingual dictionary content)
    - *Oxford Lexical Data Model* (as above)
    - *Oxford Links Data Model*, for cross-dataset links.

Dictionary datasets and their corresponding *sense-link annotations* will be used as input to the components that will be developed here. *Data schemas* are the data models supporting both input and output datasets.

3.5.3 Requirements

Functional Requirements

In a previous project we developed a sense linker, system for identifying sense links between ODE and the English side of any of our bilingual dictionaries. Pilot II.A will build on top of that, addressing in particular the following aspects:

1. **Estimate the quality of the automatically classified sense links** that are generated by our sense linking system, to avoid the need for manual validation. The goal is to be able to filter out sense pairs with low degree of reliability so we could increase the precision of the system at the cost of the recall.
   This method can also help identify which subset of links are likely to be misclassified (wrongly classified links/missed links), so that subsequent manual post-processing, such as humans carrying a peer review process, can focus on smaller subsets of data, hence reducing manual costs.
    - *Input:* A dataset of sense links automatically generated by the OUP sense linker system, in a non-LD format.
    - *Output:* (a) A dataset of sense links in the same non-LD format as the input, and (b) metadata representing how this dataset has been generated (source dataset, quality parameters applied, etc.)
2. **Classify different types of sense links based on differences in meaning granularity.**
   When aligning (or linking) the senses for a lexeme in one dictionary with the senses for the same lexeme in another dictionary, the following can be observed:
    1. Sense S_A (from dictionary A) fully aligns with sense S_B (from dictionary B). In other words, the definitions in each dictionary refer exactly to the same meaning.
    2. There is no perfect alignment between sense S_A and sense S_B because (at least) one of the definitions extends beyond the meaning conveyed by the other one.
    3. Sense S_A does not align with any sense in dictionary B.

   A system sensitive to these differences is key for ensuring the quality of any further content to be derived from dictionary-to-dictionary and dictionary-to-corpus sense links, e.g., new bilingual dictionaries, multilingual wordlists, example sentences, etc. For example, enriching content in one dictionary with complementary information from another (see the first internal use case above) can only be safely done when the senses in both dictionaries have the same meaning extent. The same situation applies to all other use cases.

- *Input:* A dataset of sense links automatically generated by the OUP sense linker system, in a non-LD format.
- *Output:* A dataset of sense links classified by sense granularity type (or degree), in a non-LD format.

These two functional requirements will be developed in parallel as independent components.

Non-functional Requirements

The following should be addressed in order to fulfill the proposed functional requirements:

- **Data integrity**. We will need to develop data models for the new types of data to produce in this pilot, that is: (a) metadata information on the output of the quality estimator system, and (b) sense link granularity annotations.
- **Interoperability: Dataset format conversion**. Datasets to be used in this pilot (ingested as input or generated as output) are not in a Linked Data (LD) format. They will need to be converted to a LD format, either in-house by extending OUP Dictionary Conversion Pipeline in order to cope with the target format or, preferably, by benefiting from the technology developed for that purpose as part of Task T3.1 of this project (*Transforming language resources and language data*), if already in place.
  Similarly, it is necessary to ensure the compatibility of sense links datasets in LD format (e.g., those generated from Task T3.2 (*Linking conceptual and lexical data for language services*)) with the components to be developed here so that the former can also benefit from the latter. Thus, the conversion from LD formats to non-LD formats should also be ensured.
  In particular, the following will need to be guaranteed:
    - **Conversion from non-LD format to LD format:**
        - Sense link datasets generated from OUP sense linker or resulting from the quality estimator system.
        - OUP monolingual dictionary formats (OxMono, LeXML, etc.).
        - OUP bilingual dictionary formats (OxBil, LeXML, etc.).
        - Metadata information resulting from the quality estimator system (see Output for Functional Requirement 1 above).
        - Sense link granularity datasets, resulting from the sense granularity classifier (see Output for Functional Requirement 2 above).
    - **Conversion from LD format to non-LD format:**
        - Sense links datasets generated from Task T3.2 of this project (*Linking conceptual and lexical data for language services),* so that it can be used as input to the two components developed in this pilot, i.e., the quality estimation system and the sense granularity classification system.
- **Reusability**. We will **take advantage of the sense linking system** already developed and available at OUP for linking the *Oxford Dictionary of English* (ODE) with the English side of OUP bilinguals (involving languages Chinese, German, French, Italian, Portuguese, Spanish).
- **Reusability: Benefitting from additional lexical linking technology.** This sub-pilot is related to the activities that will be carried out under Task T3.2 (*Linking conceptual and lexical data for language services),* and thus we will explore possible areas of collaboration with that part of the project to mutually benefit each processes and outcome. For example, results from our sense linking system can be used as baseline for that task. Similarly, results from the system to be deployed there should be able to benefit from the quality estimation system and/or the sense granularity

classifier.

- **Integrability.** The generated components will be **integrated as part of a pipeline** together with the sense linking system already developed at OUP.
- **Parametrization.** The quality estimator system will accept different **configuration parameters** on the expected levels of quality to attain.
- **Testability.** Sense link granularity annotations automatically generated from this pilot will be **evaluated against a manually annotated gold standard**, created for that purpose. That may also require the development of a simple annotation tool.
- **Testability.** The quality estimator and sense granularity classifier systems will be also **evaluated against a baseline**. In particular for the quality estimation, the baseline will consist of a single probability threshold of 0.5. Any link with a probability above that threshold is kept whereas any link below 0.5 will be disregarded. As for the sense granularity classifier, the baseline will assume that all links belong to a single “equivalent” class indicating that there are no differences in meaning between the two linked senses.
- **Compliance:** The available **license agreements** should either allow integration into the sense-linking system without requiring sense-linking system code to be made open source (e.g. Apache or MIT license) or a commercial license needs to be available.

3.5.4 Milestones

Concerning the fulfilment of functional requirement 1 (i.e., Quality Estimator):

Milestone ID	**Goal**	**Due date (Project Month)**
MS-II.1a	An approach to evaluate the levels of quality of datasets containing automatically created sense links is defined	M6
MS-II.2a	The approach is implemented	M18
MS-II.3a	Deliver draft report	M18
MS-II.4a	The approach is evaluated	M20
MS-II.5a	Final report produced	M20
MS-II.6a	Deliver final report	M36

Regarding functional requirement 2 (i.e., Sense Granularity Classifier):

Milestone ID	**Goal**	**Due date (Project Month)**
MS-II.1b	An approach for identifying different types of sense links is defined	M6

MS-II.2b	Design simple annotation tool	M6
MS-II.3b	Needed data sets are set up and annotated	M12
MS-II.4b	The approach is implemented	M18
MS-II.5b	Deliver draft report	M18
MS-II.6b	The apoproach is evaluated	M20
MS-II.7b	Final report produced	M20
MS-II.8b	Deliver final report	M36

3.5.5 Benchmarks and Evaluation

Intrinsic Evaluation

Pilot II.A will be evaluated using statistical metrics. We will assess different aspects of the results on the grounds of the following measures:

For the task in functional requirement 1:

Coverage: Total number of senses in each dictionary that should have at least one link and have effectively been linked by the system.

Precision: Number of correct sense links among all the sense links returned by the system. We will assess precision for each class (senses will be classified based on part of speech, PoS), as well as micro- and macro-average.

Recall: Number of correct sense links that have been identified by the system.[4] As before, assessing it for each PoS class, and then micro- and macro-averaging the results.

Accuracy: How well the quality the estimator fits the precision and recall of the links that were automatically generated.

For the task in functional requirement 2:

Precision: Number of sense links of each sense granularity type correctly classified. We will assess precision for each class, as well as micro- and macro-average.

Recall: Number of sense links of each type identified by the system. Assessing it for each sense granularity class, and then micro- and macro-averaging the results.

Kappa: Cohen Kappa to assess inter-annotation agreement between the system and human annotations. This same metric will have been applied to evaluate the quality of the manual annotations, which will give us an approximation on the complexity of the task.

[4] Note that this is different than coverage because the latter focuses on senses, whereas here the unit is sense links. Take for example a sense that can participate in 2 sense links. If the system is only able to return 1, this will penalize recall but not coverage.

Using the metrics above, we will compare the results obtained by the system against:

(a) **A baseline** created for that purpose for each component as defined in the non-functional requirements section.

(b) **The manual annotations (gold standard)** were generated for training the components, applying X-fold cross validation.
In the case of the component for estimating sense link quality (Functional requirement 1), the gold standard corresponds to the annotations used for training the already available OUP sense linking system. In the case of the component for identifying types of links depending on differences in sense granularity (Functional requirement 2), the gold standard will be created as part of the sub-pilot.

Contribution to KPIs

Pilot II.A will be able to directly contribute to KPIs:

- 4.1 (*Increase efficiency/speed-up adaptation to new domains and languages (development time / time to market*))
- 4.2 (*Average reduction in the time (man hours) to convert a resource to linked data at end of project*), and
- 4.3 (*Increase of stakeholder cost savings/sales as a result of Prêt-à-LLOD technologies and solutions*).

We will assess differences in costs, efficiency and time with the system set in place here in comparison to our benchmarks presented in the table in section KPIs above.

Evaluation Based on Other Standards

At the end of each task, we expect the technologies developed to reach a TRL Level of 7.

3.6 Pilot II.B: Linking corpora to dictionaries at the level of meaning

3.6.1 Use Cases Enablers

Similarly to the previous sub-pilot II.A, the link data resulting from this pilot is considered a crucial step towards the development of corpus derived products. It can be productively used to generate new content or enrich existing data. We therefore take the use cases defined next as enablers. These are:

Type	Use case	Example
Internal	Enrich current dictionary content by complementary information from corpus.	1. Include more example sentences per sense. 2. Enrich data for a specific domain, e.g. expand definition of *bottom line* by including finance definition: *Bottom Line* is the total amount a business has earned or lost at the end of the month. 3. Identify corpora novel senses for words already present in a dictionary, or senses that in spite of

		not being new, are not included yet in that dictionary.
Internal/External	Develop new linguistic data products	1. Word usage/trend information: ● word frequency at the sense level ● Collocates information at the sense level 2. Sense-tagged corpora for training and developing Natural Language Processing (NLP) applications, such as: ● Word sense disambiguation systems ● Semantic role labelling systems (which support machine translation applications) ● Systems relying on discourse structure analysis 3. Train domain- or genre-specific NLP systems, e.g. for developing social media tools
External	Support filtering and search functions in document retrieval or data mining	1. Look for the most frequent senses as an indicator for: ● guiding a search ● filtering results by most frequent sense first 2. Provide domain or genre specific wordlists, e.g. extract only medical senses of common words. 3. Link phrases/chunks to other corpus resources to compare and enrich annotations.

3.6.2 Data Sources

This pilot deals with links across two different types of datasets (dictionaries and corpora) but in the same language, which will be English and, possibly, Hindi (time allowing and if the needed resources are available).

- **Dictionary datasets**
  - *Oxford Dictionary of English* (ODE)
  - (TBC) A Hindi monolingual dictionary

- **Corpus content:**
  - *Oxford English Corpus* (OEC):
    Corpus of English containing 2.5B tokens. Covering the decade of the 2000s and with content classified by timestamp, subject and region. Data source selection and classification was highly curated and so the quality of the content is quite high.
  - *New Monitor Corpus* (NMC):
    Monitor corpus of English which contains 9.5B tokens. As before, it covers the decade of the 2000s, with content updated on a monthly basis. Documents are classified by timestamp, domain, and region.
  - (TBC) A Hindi corpus

- **Data schemas**, OUP proprietary
  - *OxMono* (for dictionary content)
  - *LeXML* (for dictionary content)
  - *Oxford Corpus Data Model* (for corpus content). Currently under development.
  - *Oxford Links Data Model* (for cross-dataset links).

Dictionary datasets and *corpus content* will be passed as input to the corpus-to-dictionary linking system that will be developed in this sub-pilot. *Data schemas* are the data models supporting both input and output datasets.

3.6.3 Sub-pilot Requirements

Functional Requirements

There is only one functional requirement for this sub-pilot, namely, align corpus data to a dictionary at the sense level. In other words, **sense-tag corpus data with the senses pre-determined in a dictionary**. The dictionary chosen for the task will have to be (a) in the same language as the source content to avoid the typical cross-lingual sense misalignments in translations, and (b) monolingual, since these tend to be informatively richer than bilingual ones (e.g., contain definitions, provide a number of example sentences, etc.).

- *Input:* (a) Corpus data, with text annotated at different levels of linguistic information. (b) Monolingual dictionary data with the sense information for the lexical elements selected for the experiment.
- *Output:* Corpus data where text has been added a layer of word sense annotations.

Non-functional Requirements

- **Data integrity. Develop corpus data model.** OUP current data model for corpus data is under development and so it will need to be completed according to the needs of this sub-pilot.
- **Interoperability: Dictionary and corpus content conversion**. As in the previous sub-pilot, dictionary and corpus sources will need to be converted to an LD format. The dictionary will have already been converted for pilot II.A. As for corpus content, a component will need to be developed for that purpose, possibly taking advantage of the technology developed as part of task T3.1.
- **Reusability. Reusing technology of OUP bilingual-to-monolingual sense linking system**. Although that system was developed for linking data of different nature, we plan to re-use the overall pipeline architecture to set the system that needs to be deployed here.
- **Effectiveness. Sampling of target lexical data** to which focus for the sub-pilot. In collaboration with lexicographers, we will identify a set of lexical items of different lexical categories over which to experiment.
- **Testability. Manual annotation of corpus content with sense tags**. We plan to use supervised machine learning methods, and therefore it will be necessary to create a training/test dataset (gold standard). This requirement will most probably involve the building of an annotation tool to simplify the tasks to annotators.
- **Testability.** The system will be also be **evaluated against a baseline**, to be defined

for that purpose**.**

- **Integrability. Collaboration with Pilot I (task 4.2),** which will address the task of corpus sense tagging but from an unsupervised approach, that is, as a sense induction task. The work carried out there can extend our system for the specific use case of identifying new senses of words already available in a dictionary.
- **Compliance.** The available **license agreements** should either allow integration into the sense-linking system without requiring sense-linking system code to be made open source (e.g. Apache or MIT license) or a commercial license needs to be available.

3.6.4 Milestones

Milestone ID	**Goal**	**Due date (Project Month)**
MS-II.1c	An approach for sense-tagging corpus data is defined	M20
MS-II.2c	Needed data sets are set up and annotated	M26
MS-II.3c	The approach is implemented	M34
MS-II.4c	The approach is evaluated	M36
MS-II.5c	Final report produced	M36
MS-II.6c	Deliver final report	M36

3.6.5 Benchmarks and Evaluation

Intrinsic evaluation

Pilot II.B will be evaluated using statistical metrics for assessing different aspects of the results.

Coverage: Total number of senses in the corpus that have been (correctly) linked by the dictionary, out of the total number of senses linked in the gold standard dataset.

Precision: Number of true sense links among all the sense links returned by the system

Recall: Number of true sense links that have been identified by the system.

Kappa: Cohen Kappa to assess inter-annotation agreement between the system and human annotations. This same metric will have been applied to evaluate the quality of the manual annotations, which will give us an approximation on the complexity of the task.

Using the metrics above, we will compare the results obtained by the system against:

(a) **A baseline** created for that purpose, namely a basic system that links every word in the corpus to the most common sense in the dictionary.

(b) **The manually annotated dataset (gold standard)** that was generated for training the components, applying X-fold cross validation.

Evaluation based on use case

We will evaluate sub-pilot II.B against an internal use case: *Enrich current dictionary content by complementary information from corpus*, in particular for updating sense data with additional example sentences.

The corpus sentences that will have been automatically identified for each sense of the targeted lexical items will be extracted, sampled, and passed to the lexicographers team, who will evaluate:

- Whether they are correctly sense-tagged.
- If positive, whether they can be directly used as dictionary example sentences or, by contrast, need to undergo some sort of manual modification (e.g., removing subordinate clauses, simplifying vocabulary, substituting offensive words, etc.).

The analysis will have both quantitative and qualitative components and will try to determine whether the feasibility of the use case is increased by means of the links support.

Contribution to KPIs

The same as for pilot II.A applies here.

Evaluation based on other standards

At the end of the task, we expect the technologies developed to reach a TRL Level of 7.

4 Pilot III: Supporting the development of cross-border public services in open government (DLX)

4.1 Overview

Derilinx supports the development of public services in open government, including cross-border services. Our Linked and Open Data Platform ***datAdore*** helps users to discover, understand and access data for their own purposes. This drives open government through evidence-based decision making,
rich business intelligence, and advanced innovation. Similarly, our public services Virtual Assistant ***GovAssist*** allows users to find information about what services are available and how to access them in a conversational manner.
The goal of this pilot is to improve the quality and accuracy of the information returned from both datAdore and GovAssist. We have therefore split the pilot into two sub-pilots
In the case of datAdore, users traditionally find data by entering keyword searches or filtering through metadata values. In this project, we will harness Prêt-à-LLOD capabilities, such as term extraction and concept disambiguation, to support unstructured, human-readable queries to cross-border data available in the platform. This facilitates the portability and uptake of cross-border urban solutions
GovAssist currently provides information about Irish public services through English. Using Prêt-à-LLOD language technologies, in this pilot we will facilitate access to information in multiple languages. For example, we will explore the possibility of providing Virtual Assistant services in the Irish language, and also in Spanish. This facilitates access to public services to a wider range of people.

Pilot III.A

Improving the quality of responses and the multilingual capability of GovAssist

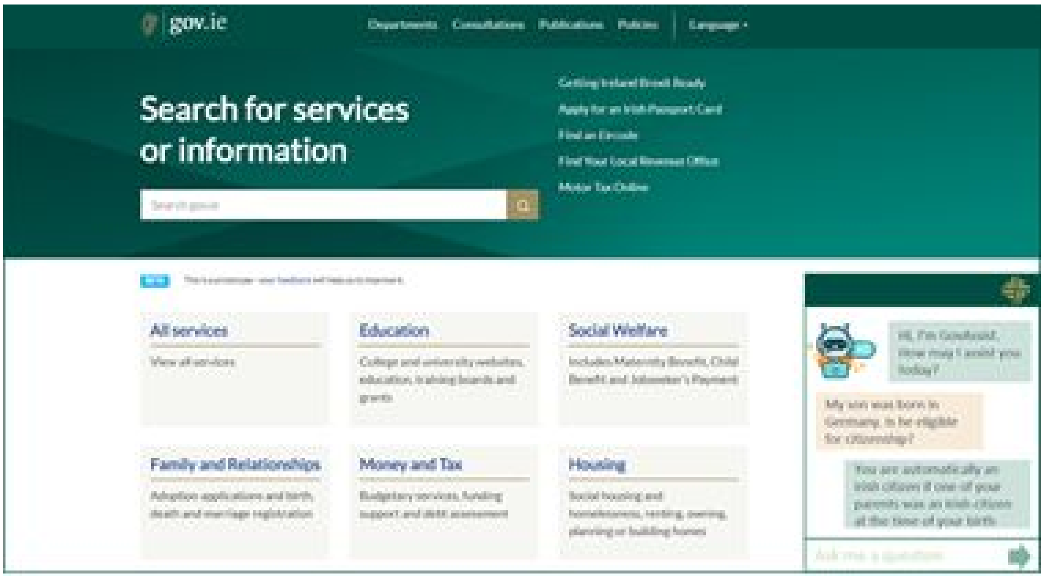
Pilot III.B

Facilitating Open Data queries on cross-border public services.

4.2 Pilot III.A: Improving the quality of responses and the multilingual capability of GovAssist

4.2.1 Objectives

In this sub-pilot, we will improve both the quality of responses and the multilingual capability of GovAssist. We will explore the possibility of providing Virtual Assistant services in the Irish language, and also in Spanish. GovAssist currently has a number of use cases. For example, together with the Irish Government's Office of the Chief Information Officer, Derilinx has developed an innovative and user-friendly chatbot that helps a wide range of people access government digital services. We focused on the 'Registers of Births, Marriages and Deaths' service, information about which comes from a number of sources including gov.ie, citizensinformation, etc. Additionally, the Irish Health Service currently operates a manual 'HSE Live' service, providing help to the Irish public in navigating the Irish public health system. They would like to enhance this by means of a virtual assistant. This chatbot will in particular improve access to the Irish Health Service's schemes and allowances programme. Prêt-à-LLOD capabilities will be used to enhance GovAssist to provide information in multiple languages, and also the quality of the responses from the chatbot.


gov.ie
Departments
Consultations
Publications
Policies
Language
Search for services or information
Getting Ireland Brexit Ready
Apply for an Irish Passport Card
Find an Eircode
Find Your Local Revenue Office
Motor Tax Online
Search gov.ie
NEW

This is a prototype - your feedback will help us to improve.

All services
View all services

Education
College and university websites, education, training boards and grants

Social Welfare
Includes Maternity Benefit, Child Benefit and Jobseeker's Payment

Family and Relationships
Adoption applications and birth, death and marriage registration

Money and Tax
Budgetary services, funding support and debt assessment

Housing
Social housing and homelessness, renting, owning, planning or building homes

Hi, I'm GovAssist, How may I assist you today?

My son was born in Germany. Is he eligible for citizenship?

You are automatically an Irish citizen if one of your parents was an Irish citizen at the time of your birth.

Ask me a question

Figure: Mock-up of GovAssist for Irish Public Services

The goal of this pilot is to provide improved access to Government services

- Chatbot enables users to interact in a conversational way and find out information about the services they are interested in.
- In many cases, users may not know which service they require, or what public body provides that service.
- Possible cross-organisation services
- Chatbot traverses these issues by guiding users to the service they require by engaging in a friendly conversation.

4.2.2 Use Cases/User Stories

Use cases that we will look at as part of this pilot are:

- How to apply for a passport
- Registering a birth
- Getting married
- Understanding benefits, schemes and allowances

Actor	Goal	Extended Goal
General user	Understand details of the public services available to me	I can access and use that public service, for example, what benefits, schemes and allowances are available to me.
Social Worker or Benefits Officer	Answer questions on the health service schemes available	I can explain to a member of the public what is available from the Irish health service

4.2.3 Data Sources

- Irish public service portal[5]
- Irish Health Services Schemes and Allowances Programme [6]
- Historical chat data from existing support systems
- Irish Health Service's Primary Care Reimbursement Service[7]
- Open Multilingual WordNet[8]
- Apertium Dictionaries[9]

Languages of interest: EN

Pre-conditions and assumptions:

- User has no technical knowledge beyond accessing website to retrieve data
- Data sources available and up to date

Tools and Technologies Involved:

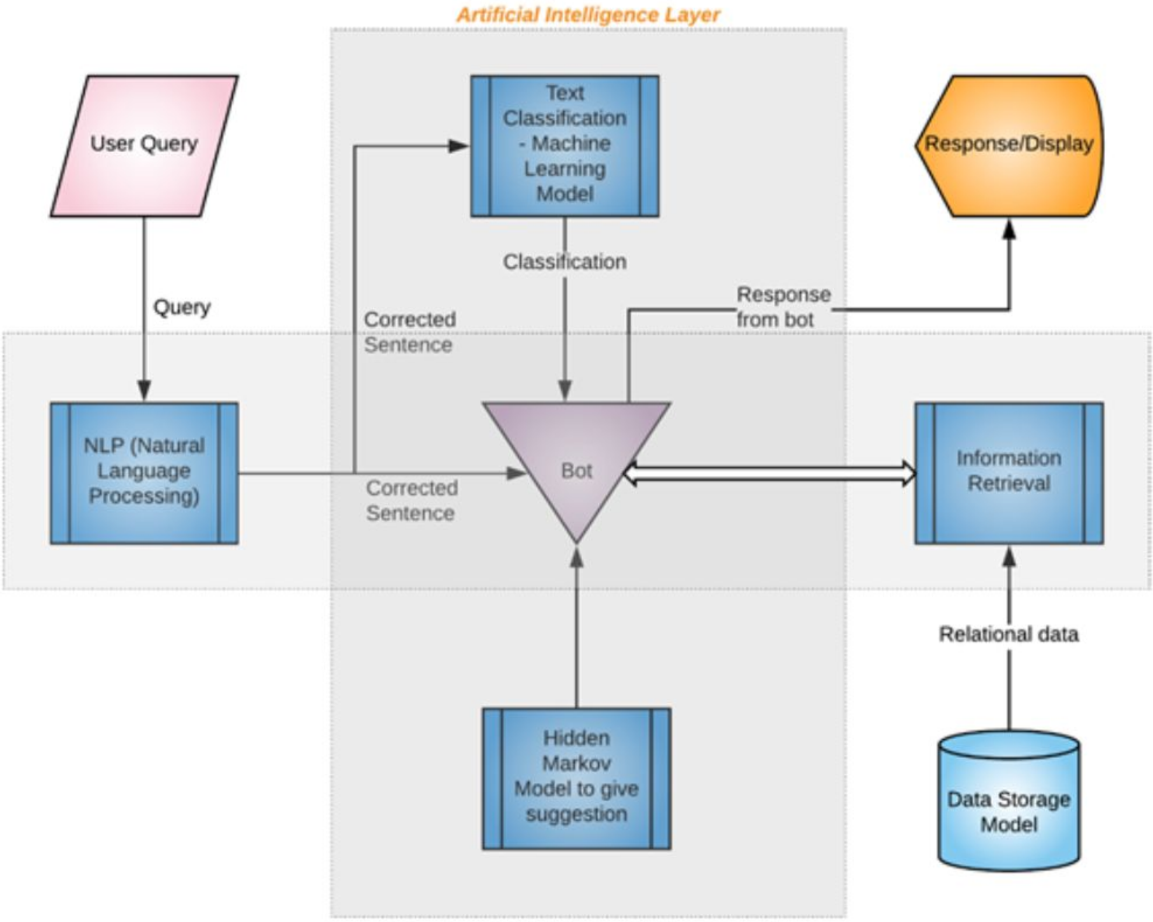
To implement GovAssist, we utilise a number of Artificial Intelligence technologies, including Natural Language Processing, Machine Learning Text Classification, Data Modelling and Information Retrieval.

- Natural language processing: NLTK (Natural Language Toolkit).
- Machine learning text classification: scikit-learn (Space Vector Model).
- Data modelling: Comma Delimited file as Knowledge base.
- Information retrieval: TF-IDF (Term frequency-inverse document frequency).

Prêt-à-LLOD capabilities to support multilingual conversations with users and to improve the accuracy of the responses.

[5] https://www.gov.ie/en/
[6] https://www2.hse.ie/costs-schemes-allowances/
[7] https://data.ehealthireland.ie/group/pcrs
[8] http://compling.hss.ntu.edu.sg/omw/
[9] http://linguistic.linkeddata.es/apertium/

Initial Overview:



Workflow

Step 1: User wants information about health services.
Step 2: User logs in to Chatbot / User skips login
Step 3: User types their (natural language) question/ AI suggests to user somewhere to start
Step 4: Chatbot passes the natural language query to the Prêt-à-LLOD tools which disambiguates it into one or many structured queries
Step 5: Chatbot displays suggested queries to user
Step 6: Chatbot asks user to select the most appropriate query or to rephrase their question
If select then Step 7
Else Step 4
Step 7: Chatbot retrieves the correct answer for the user's question using the structured query text provided by Prêt-à-LLOD
Step 8: Chatbot checks answer against QA process (ensures answers remain accurate as bot is training)
Step 9: Chatbot displays the answer to the user.
Step 10: Chatbot asks user if question has been answered satisfactorily
If yes then Step 11

Else

- ask them to phrase the question differently
- would they prefer to talk to an agent
    - If yes then Step 13
- Step 4

Step 11: Chatbot gets the suggested next question

Step 12: Chatbot suggests the next question to the user

If the user selects the suggested question or types their own question, repeat from Step 4

If the user says goodbye or similar, say goodbye and close.

Step 13: Chatbot transfers to agent

Post conditions:

- New iteration of user-suggested question/structured query pairs and answers which are used to improve the model
- User's question has been answered to their satisfaction
- Some measure of the consistency of responses collated

Exceptions/errors:

- The user's question is not clear or is irrelevant

4.2.4 Requirements

Functional Requirements

Pilot III.A will enhance the GovAssist Chatbot, focussing in particular on:

- The availability of the chatbot in multiple languages, possibly including the Irish language.
- Improving the interpretation and disambiguation of natural language questions
- Enhancement of the GovAssist AI model
- Transfer to agent function (on request from user)
- QA system to provide consistency of answers and ability to audit answers

Non-Functional Requirements

The following should be addressed in order to fulfill the proposed functional requirements:

- The GovAssist Chatbot should be extensible to facilitate the incorporation of Prêt-à-LLOD functionality as it becomes available and evolves
- The Chatbot must comply with GDPR and ethical requirements
- The available license agreements should either allow integration into the Chatbot without requiring Chatbot code to be made open source (e.g. Apache or MIT license) or a commercial license needs to be available.

4.2.5 Milestones

Milestone ID	Goal	Due date (Project Month)
MS-III.1a	High level design agreed	M8
MS-III.2a	Development of test cases for internal (DLX) testing	M10
MS-III.3a	Alpha version of application ready for testing	M12
MS-III.4a	Beta version of application ready for testing	M18
MS-III.5a	Application completes internal testing	M20
MS-III.6a	Pilot user satisfaction survey developed	M21
MS-III.7a	User testing complete	M22
MS-III.8a	Application fully implemented	M24

4.2.6 Deliverables

Pilot III.A deliverables will consist of the following:

- **Irish Health Service chatbot** incorporating Prêt-à-LLOD functionality as detailed above
- **Technical documentation** generated as part of the development project
- **Contribution to KPIs**
- **Final report**, briefing on the development activities and reporting the results of the project

4.2.7 Benchmarks and Evaluation

Application Benchmarks

- Quality of response:
  - Analysis of content of application responses against agent responses, tested using historic live chat data supplied by HSE

- Beta Testing: User testing with trial users (internal to HSE) incorporating:
  - anonymous and named use of application
  - analysis of QA data for consistency metric
  - functionality of transfer to agent function
  - satisfaction of agents with transfer to agent functionality

- Metrics for Application Usage:
    - percentage of queries transferred to agent
    - number of customers
    - number of queries addressed
    - percentage of queries satisfactorily addressed

KPI Contributions

Nr	KPI		Benchmarks
4.4	Satisfaction level of early adopters of pilots in a survey	4-5/5	Chatbot pilot user satisfaction survey

4.3 Pilot III.B: Facilitating Open Data queries on cross-border public services

4.3.1 Objectives

A member of the public would like to discover information about a service in a cross-border context, on the European Open Data portal or across a number of different national Open Data portals, through browsing a catalogue of the datasets.This data may be in any European language (metadata, data dictionary). The discovered data will be displayed to the user in their native language; as a minimum this includes the metadata but preferably the data dictionary as well. The user can then access, download and share the discovered data. The goal of the pilot is to reduce costs and time spent by the general public in accessing cross-border data and improve access to this data.

4.3.2 Use Cases/User Stories

Actor	Goal	Extended Goal
Tourist	access open transport data on my destination	I can plan my journey
Teacher	access open data from across Europe	I can teach my pupils about other countries
Journalist	access open data from other countries	I can write articles comparing life in my country with those countries
Home Buyer	access open data on my new location	I know what to expect when I move house

4.3.3 Data Sources

- European data portal (https://www.europeandataportal.eu/)
- National Open Data portal of specified European countries
- Open Multilingual WordNet
- Apertium Dictionaries

Languages of interest: EN, RO, SI, FI

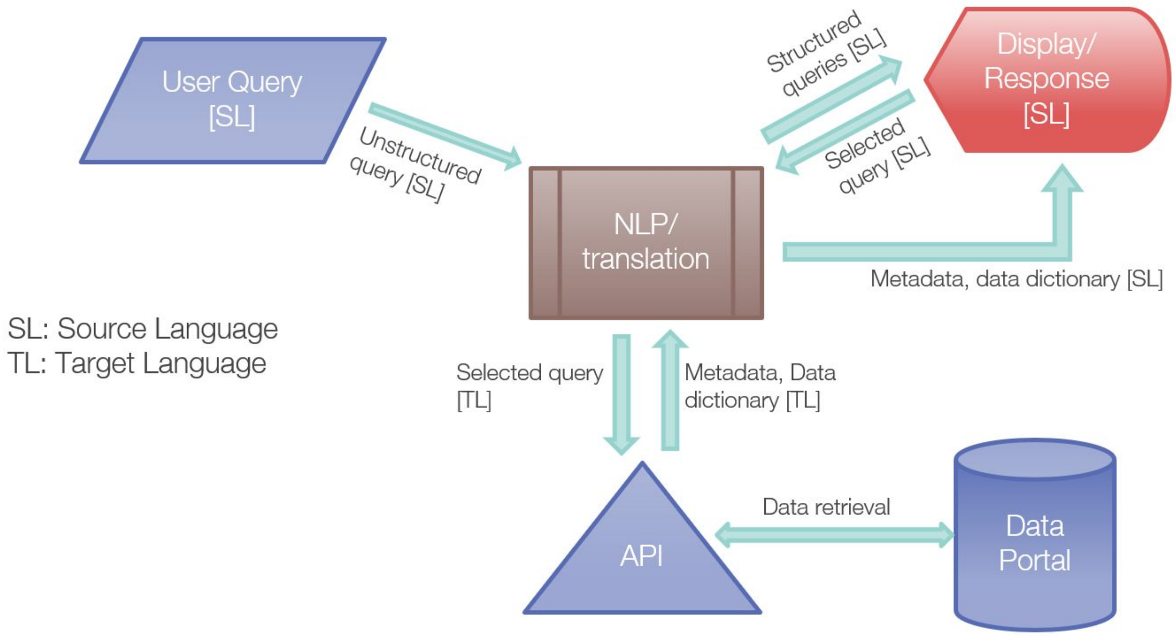
Pre-conditions and assumptions:

- Relevant data has previously been published on the relevant Open Data portal
- The user does not have knowledge of all languages
- The user has no technical knowledge beyond accessing portal to retrieve data
- The language resources required are supported by Prêt-à-LLOD

Tools and Technologies Involved:

- datAdore Linked and Open Data PLatform
- architecture: Data Model + Natural Language Model + Machine Learning Model + Information Retrieval + Hidden Markov Model (Postgres + Python)
- GUI: HTML + Bootstrap + Javascript
- Prêt-à-LLOD capabilities for the discovery of relevant language resources and linking of resources across data sources.

Overview:



Workflow:

Step 1: A user has a cross-border public-service query
Step 2: The user specifies through datAdore or the API the language they wish to work in ('source language')
Step 3: The user poses a query about a public-service in natural language in the source language.
Step 4: By harnessing Prêt-à-LLOD capabilities, such as the discovery of relevant language resources, the query is then broken down and executed across the Open Data sources.
Step 5: The dynamic query results are presented to the user in their own language, along with links to the source data.
Step 6: The Query GUI Tool displays the selected data. The user can then choose to share or download the data

- **If Share:** The user will be given a generated URL to share, which when dereferenced, will display the data the user is currently viewing
- **If Download:** The user will be given the option to download the data

Post-condition:

- Cross-border data has been shared via a browsable URL
- Cross-border data has been downloaded

Exceptions/errors:

Relevant data is not available from the requested countries

4.3.4 Requirements

Functional Requirements

This pilot will focus in particular on:

- Interpretation and disambiguation of natural language queries
- Development of an API that enables cross-border Open Data discovery through selected (European) languages
- Extension of datAdore based on the API, to allow the user to identify data relevant to their search through selected (European) languages
- Translation assist tool to allow translation of metadata and data dictionary into selected (European) languages
- The API will work in conjunction with the European data portal and other Open Data portals within Europe

Non-functional Requirements

The following should be addressed in order to fulfill the proposed functional requirements:

- The extension of datAdore should allow the incorporation of appropriate Prêt-à-LLOD workflows as it becomes available and evolves
- The available licence agreements should either allow integration into the Tool without requiring code to be made open source (e.g. Apache or MIT license) or a commercial license needs to be available.

4.3.5 Milestones

Milestone ID	**Goal**	**Due date (Project Month)**
MS-III.1b	High level design agreed	M12
MS-III.2b	Development of test cases for internal (DLX) testing	M18
MS-III.3b	Alpha version of application ready for testing	M24
MS-III.4b	Beta version of application ready for testing	M30
MS-III.5b	Application completes internal testing	M32
MS-III.6b	User testing complete	M34
MS-III.7b	Application fully implemented	M36

4.3.6 Deliverables

Pilot III.B deliverables will consist of the following:

- **Application providing improved access to cross-border Open Data** as detailed above
- **Technical documentation** generated as part of the development project
- **Contribution to KPIs**
- **Final report**, briefing on the development activities in each of the sub-pilots and reporting the results of the project

4.3.7 Benchmarks and Evaluation

Application Benchmarks

We will evaluate the sub-pilot for the quality and speed of its results against a manual process.

The time and quality benchmarks will be set by a user with a translation tool, measuring the time taken to identify and retrieve relevant data.
The datasets that have been identified by the application will be marked for comparative relevance and the time taken to identify and retrieve.

The criteria:

1. Were all the datasets manually identified as relevant identified by the application?
2. How relevant were any additional datasets that were retrieved using the application?
3. Was the data presented to the user in a meaningful form?
4. What was the time saving in delivering meaningful results with the application vs. the manual process?

The analysis will have both quantitative and qualitative components.

- Beta Testing: User testing with trial users incorporating:
    - Criteria 1,2 and 3 above
    - Satisfaction of trial users with functionality

- Metrics for Application Usage:
    - Number of datasets downloaded using the tool
    - Number of queries made using the tool

KPI Contributions

Nr	KPI		Benchmarks
4.2	Average reduction in the time (man hours) to convert a resource to linked data at end of project.	90%	Develop benchmarks based on sample input from Slovenian, Romanian, Finnish users - manual method vs with application

4.4	Satisfaction level of early adopters of pilots in a survey	4-5/5	Survey of Slovenian, Romanian, Finnish users

5 Pilot IV: Multilingual Text Analytics for Extracting Real-World Evidence in the Pharma Sector (SEM)

5.1 Objectives

The goal of this pilot is to bring flexible multilingual capabilities to the Semalytix Pharos® platform, a tool for pharma companies to analyse real-world evidence (RWE). In the context of pharma, RWE means evidence extracted from data other than controlled clinical studies in order to reveal insights about the performance and patient outcomes of a treatment under real-world conditions, outside the lab. To enable customers to derive meaningful insights from RWE, the Pharos® plattform revolves around individual collections of data visualizations (dashboards) which are designed to answer specific analytical questions. Each of these dashboards is based on a particular selection and orchestration of analytical components from the Semalytix technology stack which includes a pharma-specific knowledge graph and ontologies, domain-specific lexico-syntactic rules and (un)supervised machine learning models. All of these components are currently optimized to process English data only.

To address questions about non-English text data, a straightforward approach would be to recreate each analytical component for that language, requiring annotators and language engineers to have knowledge of the language and the pharma domain, so that they can annotate data, create ontologies and lexico-syntactic rules, as well as generating meaningful feature representations for machine learning models. As this is both time-consuming and costly, we seek to minimize the need for manual efforts by using LLOD resources to enable language transfer of existing systems. Given the diversity of components used in a single dashboard, one-off transfer for a specific NLP approach will not suffice. Rather, depending on the type of the respective component to be transfered, language transfer will need to involve different approaches (“recipes”) and specific resources -- ranging from parallel corpora which help train task-specific cross-lingual embeddings to multilingual linked data for bootstrapping dictionaries for entity tagging. Therefore, our goal is to develop a framework for configurable language transfer pipelines enabled by the capabilities to discover, transform and compose language resources developed within the Prêt-à-LLOD project.

As the current questions and use cases targeted with Pharos® are primarily from the pharma sector and involve domain-specific terminology and data, language transfer also implies the need for domain preservation depending on the domain characteristics of the used multilingual resources. We expect openly available resources to be mostly domain-agnostic (hence, not specialized for pharma) and will develop domain adaptation strategies to mitigate this gap. This will lead to a generic methodology for adaptation to other domains and will enable us to perform first tests on domains which as of now have not been addressed by Semalytix, such as finance, with an extended set of languages.

5.2 Use Cases/User Stories

As outlined in 5.1, Semalytix implements an overarching user story within Pilot IV: “As a language technology services provider, we want to accelerate customer-facing time-to-deployment for domain-specific text analytics services in multiple languages in order to save internal resources and scale service provision capabilities”. This is broken down into a number of user stories which describe the steps and roles involved in the creation of language transfer pipelines. As discussed above, these pipelines will need to support transfer for different types of approaches: ML models of different kinds, lexico-syntactic rules and lexical resources/ontologies.

Actor	Goal	Extended Goal
NLP architect	specify approaches for cross-lingual transfer of different types of NLP systems	language resources required for each transfer approach can be determined
NLP architect	specify approaches for domain adaptation of particular language resources	domain preservation within cross-lingual model transfer is guaranteed
NLP service engineer	specify criteria for language resources that are required by a language transfer approach	at least one recipe can be implemented for each class of NLP systems
language engineer	search for language resources according to predefined criteria (resource type, language/language pair, domain, license)	more efficient discovery
language engineer	be able to combine information from various complementary language resources	increase potential coverage of the transferred system
language engineer	receive suggestions about which language resources might be complementary and interoperable	identification of data sources which are relevant to the transfer approach is accelerated
NLP service engineer	receive language resources in a unified format	transfer algorithms can abstract from differences in data representation

language engineer	automatically convert language resources from a given input to a given output format	make different resources interoperable
NLP service engineer	implement domain adaptation approach for language resources	out-of-domain or domain-agnostic resources can still be used for cross-lingual transfer of domain-specific models
system architect	define transfer pipelines for different approaches in a declarative way	the deployment of all necessary components can be automated
forward-deployed engineer	execute defined transfer pipelines on demand for existing systems	transferred models can be used in production faster

5.3 Data Sources

5.3.1 Multilingual pharma sales-interaction corpus

The main data source used in pilot IV is a dataset containing protocols of sales and market research interactions between health care practitioners (HCP) and pharma sales representatives from the HCP's perspective. The protocol forms contain structured questions as well as fields for entering free text. HCPs are requested (and compensated) to fill out a protocol form after any interaction.

The most salient feature of these data is a free-text field filled out by HCPs to answer the question "What did you discuss today?". Its contents vary greatly, also in length, but mostly contain rather formal and condensed language, technical terms and abbreviations. This free-text field is the primary input for our NLP systems from the sales interaction corpus. In accordance with the language transfer use case, the pilot uses data in multiple languages which are English and Spanish primarily, but also include German, French and potentially Japanese.

5.3.2 Multilingual LLOD resources

Language transfer of different types of NLP systems or models requires corresponding multilingual resources for each type and transfer strategy. Thus, the pilot project will make use of a broad range of LLOD resources[10], especially as their discoverability is further improved in the context of Prêt-à-LLOD. Bilingual lexica, in particular the RDF versions of Apertium, are expected to be a key resource across all types of systems and models in

[10] http://linghub.org/

scope (machine learning models over discrete and continuous feature spaces, lexico-syntactic rules, ontologies and lexical resources). The Apertium dictionaries also exemplify the gap between large, readily-available multilingual resources which are open-domain and the requirements of language transfer for specific domains. As such, they will form a robust basis for baseline approaches as suitable techniques for domain adaptation and more in-domain resources are explored. Similarly, specialized resources for individual types of NLP systems will be evaluated and specifically created, if necessary. In particular, we will explore the use of parallel corpora and multilingual embeddings for the transfer of supervised machine learning systems. Transferring lexico-syntactic patterns might employ different resources to evaluate cross-lingual syntactic equivalences. Pharma-specific knowledge graphs could benefit from employing multilingual LOD resources, e.g. derived from Wikipedia or similar sources, to transfer lexicalizations.

5.4 Requirements

5.4.1 Functional Requirements

- **Creation, configuration and deployment of language transfer pipelines**
    - Consume LLOD resources as part of language transfer pipelines
    - Configuration of language transfer pipelines, selecting processing steps, LLOD resources (and how to transform and combine them)
    - Deployment and execution of pipelines for language transfer that interface with the specified resources
- **Discovery and wrangling of LLOD resources relevant for language transfer**
    - Search for LLOD resources according to predefined criteria (resource type, language/language pair, domain, license)
    - Combine complementary LLOD resources, in the sense of handling them within language transfer as if they were a single resource
    - Based on a given LLOD resource, suggest interoperable and complementary resources in the context of language transfer
    - Transform between different formats of LLOD resources
- **Transfer of NLP systems based on different types of approaches from a source language to a target language**
    - Transfer supervised machine learning models based on embedding, distributional, morphological or linguistic features
    - Transfer lexicalisations of subgraphs of a knowledge graph as used e.g. for entity tagging
    - Transfer patterns based on lexico-syntactical features and entity types
    - Transfer patterns based on unsupervised topic extraction

5.4.2 Non-functional Requirements

- The costs and implied person hours of employing language transfer pipelines should be less than for recreating each analytical component for the target language.
- The performance of the transferred NLP systems should be at least within a reasonable margin below the performance of target-language systems.

- The solution should be flexible in terms of configurability and support for different types of NLP approaches and resources, i.e. abstract from the implementation of individual NLP systems.
- The creation of specific language transfer pipelines should be at least semi-automatized.

5.5 Milestones

Milestone ID	**Goal**	**Due date (Project Month)**
MS-IV.1	Prototype implementation of language transfer for at least one type of component (e.g. supervised ML) and at least one pair of source and target language (English and e.g. Spanish)	M9
MS-IV.2	Individually implemented language transfer of all analytical components needed to populate a complete dashboard	M15
MS-IV.3	Discover and consume cross-lingual LLOD resources and services supporting manual language transfer of a complete dashboard	M18
MS-IV.4	Language transfer pipeline consuming LLOD resources for at least *one type of analytical component;* Pilot Report Version 1	M24
MS-IV.5	Language transfer pipelines consuming and transforming LLOD resources for *all analytical components* needed to populate a complete dashboard	M32
MS-IV.6	Evaluations complete	M33
MS-IV.7	Documentation complete	M35
MS-IV.8	Pilot Report Version 2	M36

Dependencies on deliverables from WP3 concern milestones MS-IV.3 (D3.2 Language Resource and Service Linking; M15) and MS-IV.5 (D3.3 Workflows for NLP services; M30).

5.6 Deliverables and Contributions

Pilot IV will demonstrate the combined benefits of solutions to two challenges addressed within Prêt-à-LLOD, the **discovery** of LLOD resources and **workflow composition** across heterogenous NLP services. Improved discovery enables the selection of resources with appropriate characteristics in a (semi-)automated fashion. As these resources will be interoperable as part of composable workflows, the varying resource requirements of language transfer for different types of NLP systems can be met efficiently to create flexible language transfer pipelines. All pilot outcomes and progress will be shared based on **an intermediate and a final report** (pilot report V1 and V2).

All this is meant to **reduce development-time** and time-to-market of systems for languages other than English, which is directly relevant for KPI 4.1. The pilot also enables a **stakeholder cost comparison** of Prêt-à-LLOD resources with commercial equivalents in the context of language transfer of NLP systems for pharma-related use cases (KPI 4.3). Early adopter clients will be surveyed for their **satisfaction level** to demonstrate the overall usefulness of the solution enabled by Prêt-à-LLOD outcomes (KPI 4.4).

5.7 Benchmarks and Evaluation

Our benchmarks are based on a comparison across three conditions:

1) Systems that rely on a machine translation service in order to translate input data from a target language of interest into English so that existing models trained on English source-language data can be directly applied

2) Systems created from scratch for the target language

3) Systems trained on source-language data and transferred to the target language using a language transfer pipeline

These comparisons will take the following dimensions into account:

1) System-specific, intrinsic evaluation, e.g. in terms of accuracy

2) End-to-end evaluation of relevance of transfered information wrt. Real-world business cases as addressed in a customer dashboard (e.g., in terms of precision based on relevant/non-relevant pieces of information being displayed)

3) "Time-to-platform" in terms of time needed for system development and deployment (KPI 4.1).

4) "Time-to-performance-at-k" in terms of time needed to achieve a particular level of k% system accuracy

In addition, we will perform a customer survey with early-adopter customers to determine their satisfaction level (KPI 4.4). Also, we will deliver an overall cost analysis of transferring language-specific NLP systems for our use cases and compare commercial with Prêt-à-LLOD resources (KPI 4.3).

6 Summary

The present report delivers specifications of four business pilots (in terms of industry use case scenarios, requirements and evaluation concepts) to be carried out within the Prêt-à-LLOD project. As detailed in the previous sections, these pilots will showcase the impact of the language technology methodologies and workflows to be developed in Prêt-à-LLOD on faster and more efficient development of services and applications across multiple industries.

Prêt-à-LLOD Challenge	**Addressed in Pilot(s)**
Discovery	1, 3, 4
Linking	1, 2
Transformation	2
Workflow Composition	1, 3, 4

Table 2: Overview of Prêt-à-LLOD challenges addressed in pilots

KPI according to SoW	**Description**	**Covered by Pilot(s)**
4.1	Increase in efficiency/speed-up of adaptation to new domains and languages (development time/time-to-market)	1, 2, 4
4.2	Average reduction in time (person hours) to convert a resource to linked data	2
4.3	Increase of stakeholder cost savings/sales as a result of Prêt-à-LLOD technologies and solutions due to usage of free language resources	1, 2, 3, 4
4.4	Satisfaction level of early adopters of pilot in survey	1, 3, 4

Table 3: Overview of pilot-related KPIs according to Statement of Work

In conceptualizing the business pilots, particular attention has been paid to align pilot activities with (a) main challenges addressed by Prêt-à-LLOD and (b) KPIs in order to measure project success in tangible metrics as stated in the Statement of Work. As can be seen from Tables 2 and 3 above, respectively, all challenges are covered in at least one pilot, as well as all KPIs being taken up.