Prêt-à-LLOD

D2.3 - Strategic Report on Business Plan Development v1

Author(s):
Katharine Cooney (DLX),
Eva Theodoridou (OUP),
Matthias Hartung (SEM),
Thomas Thurner (SWC),
Pierre Baviera (DLX),
Katherine Martin (OUP),
Thierry Declerck (DFKI)

Date: 3/6/2020

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

D2.3 - Strategic Report on Business Plan Development v1

Deliverable Number: D14
Dissemination Level: P(ublic)
Delivery Date: M18
Version:
Author(s): Katharine Cooney (DLX), Pierre Baviera (DLX), Matthias Hartung (SEM), Eva Theodoridou (OUP), Thomas Thurner (SWC), Katherine Martin (OUP), Thierry Declerck (DFKI)

Document History

Version Date	Changes	Authors
13/02/2020	First draft	DLX
27/4/2020	Added BMC for OUP and SWC	OUP, SWC
5/5/2020	Added BMC for SEM	SEM
11/5/2020	Added Competitor Analysis	SEM
14/5/2020	Added Market Analysis	SEM
28/5/2020	Addressed comments	DLX
8/6/2020	Formatting	DLX

Executive Summary

This deliverable presents the first version of the strategic business development plan for Prêt-à-LLOD, focussing on the industrial partners' specific business plans.

The Prêt-à-LLOD industrial partners are:

Semantic Web Company[1] (SWC)

Oxford University Press[2] (OUP)

Derilinx[3] (DLX)

Semalytix[4] (SEM)

All Prêt-à-LLOD industrial partners have been involved in the documentation of their respective business development plans and thereby in the creation of this deliverable with the aim of creating a comprehensive picture across the industrial pilots of the Prêt-à-LLOD project.

Further activities will be reported in month 30 of the project (D2.4).

The deliverable is structured as follows:

The first section introduces the report, documents this report's relationships with other Prêt-à-LLOD project deliverables and the expected impacts the pilots will have in their respective domains.

The second section describes the industrial partner companies and pilots and illustrates their business plans using the Business Model Canvas.

The third section includes market analysis for Language Technology and Natural Language Processing and competitor and market analysis for each of the pilots.

The fourth section presents a SWOT analysis in order to investigate the macro-environment and ecosystem of the Prêt-à-LLOD project.

The report concludes with a summary and the next steps planned by the partners.

[1] https://semantic-web.com/

[2] https://global.oup.com/

[3] https://derilinx.com/

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

Contents

Appendix

1. Introduction

The aim of the pilots is to demonstrate the commercial potential for application of the Prêt-à-LLOD tools and methodologies. This ensures that results will be used after the project is completed, but also that the consortium's partners will derive real benefit and added value by making use of the respective project results.

This document elicits the current business plan for each of the pilots, defining and differentiating the proposed advantages. It includes preliminary details on the revenue generation model for each pilot utilizing the business canvas method and presents conclusions as to next steps.

1.1.Background

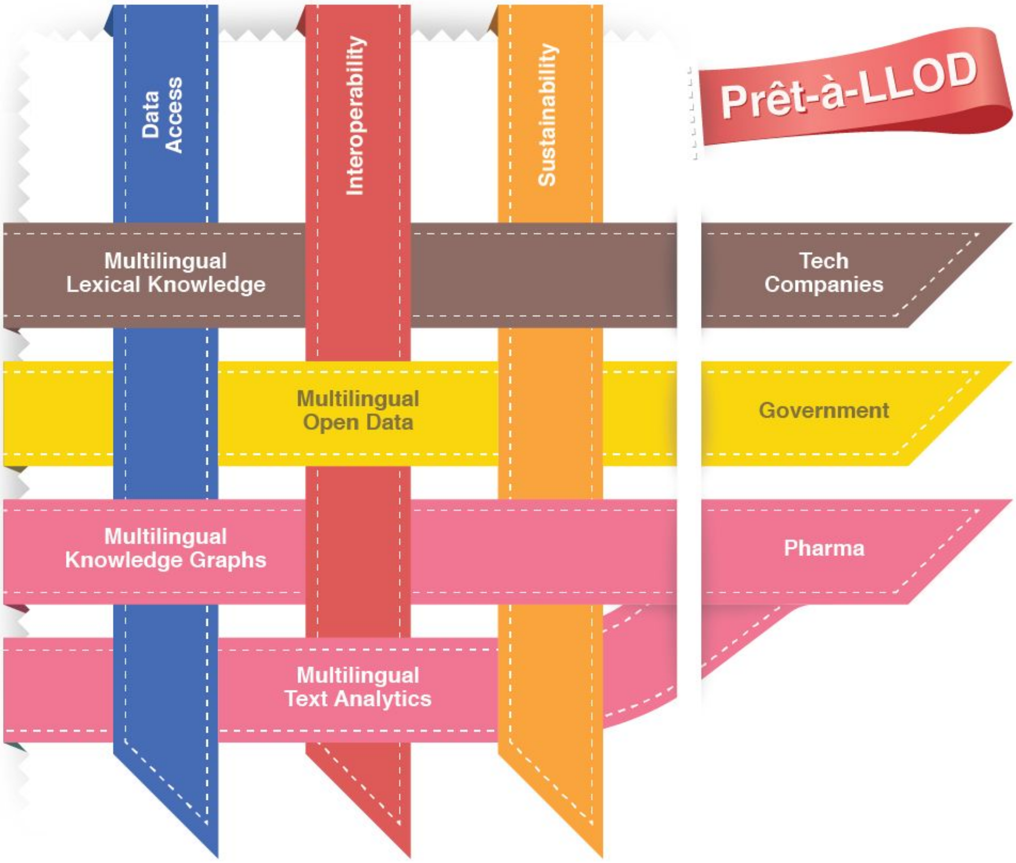


Figure 1: Overview of objectives, uses cases and sectors for Prêt-à-LLOD

This diagram depicts the interaction between the industrial pilots and the overall objectives of the Prêt-à-LLOD project. The horizontal ribbons depict the key focus areas for the pilots (multilingual lexical knowledge, multilingual Open Data, multilingual knowledge graphs and multilingual text analytics) and the commercial sectors in which the industrial partners operate (tech companies, government and pharma). The vertical ribbons depict the overall goals of the project (data access, interoperability and sustainability).

The Digital Single Market is a European Commission strategy that aims to create a next generation of the Internet, where European companies can easily sell their products and services across a market of more than 500 million people. Prêt-à-LLOD aims to create new data value chains for language resources and language services across several sectors and application areas (depicted in Figure 1) and thus contribute to the realisation of the Digital Single Market.

One of the biggest challenges of the Digital Single Market is that it is spread across many countries and languages and, as such, the ability to quickly adapt tools to new markets is of vital importance. The persistence of language barriers on the Web has been recognized by the European Commission as a genuine European issue. In this context, some EU funded reports[5] identified *language technologies*[6] and *linked data*[7] as core technologies in the roadmap for a Multilingual Digital Single Market.

Language technologies are a market sector which is predicted to grow to over $22 billion by 2025[8] and are one of the fastest growing sectors in the economy. As more sources of open data become available, applications that integrate data across a variety of sectors, including the pharmaceutical, technology and government sectors, promise significant cost savings and new commercial opportunities. The integration of data from multiple sources necessarily requires *transformation* of those data into an interoperable format. This in turn allows the *linking* of those datasets for their deployment in an effective *workflow*. This reflects the main technological aspects of Prêt-à-LLOD: Transformation, Linking and Workflow[9].

Prêt-à-LLOD's principal objective is to utilize linked open data and language technologies in order to create ground-breaking cross-sectoral applications.

[5] E.g., the Strategic Research and Innovation Agenda (SRIA) for the Multilingual Digital Single Market, commissioned to the "Cracking the Language Barrier" federation and elaborated by many experts of European projects and organisations working on multilingual technologies.

[6] By language technologies (LT) we mean Natural Language Processing (NLP) technologies

[7] Refers to a collection of interrelated datasets on the web, reachable and manageable by Semantic Web tools. See https://www.w3.org/standards/semanticweb/data.html

[8] https://www.tractica.com/newsroom/press-releases/natural-language-processing-market-to-reach-22-3-billion-by-2025/

[9] D2.3 "Research Challenge Report v2", will detail the methods and technologies

The industry partners have been selected to cover key areas of the data value chain for language resources. Firstly, **Oxford University Press** develops some of the world's most renowned language resources, including the *Oxford English Dictionary*; through its Oxford Dictionaries API[10] it provides monolingual and bilingual data in many languages to a broad range of enterprise customers, and through its language data licensing program it provides language data to major technology companies such as Google. **Semalytix** is a spin-off company from the University of Bielefeld. It specializes in machine reading and text analytics solutions for business intelligence applications which mainly serve customers from the global pharmaceutical industry. **Semantic Web Company** is a world leader in the use of linked data for metadata, search and analytic solutions. Finally, **Derilinx** provides services to government for high-quality data publishing and is a leading Linked & Open Data company, driving decision-making and providing insights in the public sector.

1.2.Relationships with other project deliverables

- D2.1 Report on User Stories
  The main goal of this task is to specify the user needs and elaborate user scenarios that will guide the design and development of the functionalities. This report is being submitted in M18.

- D2.2 Report on Community-Driven Requirements
  This report, submitted in M18, documents conversations with project stakeholder groups as requirements and use cases, grouping them as research challenges and pilot activities.

- D6.2.1 Project Exploitation Report
  This report, version 1 of which is due in M24 of the project, will report progress on the individual and collective exploitation plans (including collaborations between project partners) as well as taking into account changes in relevant market sectors and emerging relevant research. It will address the success and financial sustainability of the project.

- WP4: pilot reports, pilot demonstrations

1.3. Expected impacts

The four pilots address specific challenges in three market sectors of high commercial and/or societal impact (as depicted in Figure 1). The technology companies' sector is addressed by Oxford University Press' pilot, "Linking Lexical Knowledge to Facilitate Rapid Integration and Wider Application of Lexicographic Resources for Technology Companies". The pharmaceutical sector is addressed by Semalytix in their pilot

[10] https://developer.oxforddictionaries.com/

“Multilingual Text Analytics for Extracting Real-World Evidence in the Pharma Sector” and Semantic Web Company in “Multilingual Knowledge Graphs for Knowledge Management across Sectors”. Finally, government services are addressed by Derilinx in “Supporting the Development of Public Services in Open Government both within and across borders”.

Each of the partners will integrate the technologies developed in Prêt-à-LLOD into their products, providing proof-of concept for new applications and product features. The result of this integration will support each of the industrial partners in reducing costs and time-to-market for their products.

1.3.1. Impact on Technology companies

Software companies incorporating natural language processing into their products rely on high quality lexical resources. As these companies expand the scope of these products to support languages outside English, the relevant lexical resources become more difficult to identify, source and prepare. Making data truly interoperable will provide these technology companies with data that can be plugged into their systems quickly and easily, meaning cost savings in data identification, sourcing and preparation.

Oxford University Press (OUP) has a strong background in lexical data and language technology. By participating in Prêt-à-LLOD, OUP aims to improve interoperability and flexibility of its data and further enhance its link creation and verification capability, so as to make data available for wider computational use and enable the creation of new language data products.

1.3.2. Impact on Pharma

The costs of regulatory compliance and drug approval are high. As part of the process, pharmaceutical companies are increasingly expected to include not only the clinical studies data, but multiple heterogeneous datasets which may be in a variety of languages. These sources could include patient records, medical and insurance claims, social media etc.

According to IDC estimates, up to 80% of healthcare data is multilingual and unstructured[11] and with estimates that the volume of healthcare data will grow at an annual rate of 36% between 2018 and 2025[12], Artificial Intelligence and Machine Learning solutions will increasingly be required to analyse this volume of data.

[11] https://www.healthdataarchiver.com/health-data-volumes-skyrocket-legacy-data-archives-rise-hie/

[12] https://www.seagate.com/files/www-content/our-story/trends/files/idc-seagate-dataage-whitepaper.pdf

Semalytix' customers are pharmaceutical companies operating in global markets; a large portion of these companies' revenue is generated in non-English markets. Semalytix' main product is Pharos® Pharma Analytics[13], a service platform that extracts insights from pharma-related data sources. Semalytix will focus on extending this service to non-English markets, in particular to German, Spanish and French.

PoolParty[14], Semantic Web Company's flagship product, provides Semantic Middleware and is designed to be applicable to multiple domains, including pharma. Through their participation in the Prêt-à-LLOD project, they aim to improve term extraction and concept matching services workflows currently existing in PoolParty, as well as extending to several new languages.

1.3.3. Impact on Health Information and Government Services

Access to healthcare and government services can be complicated, even more so for the users trying to access benefits through a language that is not their own. To facilitate easier access to this information, AI/NLP systems can be used to complement other communication channels. Conversational User Interfaces (CUI) such as chatbots can provide more uniform (and multilingual) access to health information and government services and have the potential to optimize citizen interactions. This can give providers the ability to increase customer service levels without increasing incremental cost to serve. Derilinx will use their pilot to provide cross-border and multilingual access to health information and government services.

2. Pilots and Business Model Canvas

1.1.Industrial partners' pilots

In Pilot I, "Multilingual Knowledge Graphs for Knowledge Management Across Sectors", Semantic Web Company aims to improve term extraction and concept matching services as offered by their flagship product, PoolParty. Three sub-pilots will look at quality enhancement in various areas as well as the replacement of certain proprietary language resources currently used with open source language resources.

The title of Pilot II is "Linking lexical knowledge to facilitate rapid integration and wider application of lexicographic resources for technology companies". Oxford University Press, as a provider of highly curated, comprehensive and lexically rich resources for the language technology industry, are using this pilot to implement a novel methodology for generating bilingual dictionaries.

[13] https://www.semalytix.com/solutions/

[14] https://www.poolparty.biz/

In Pilot III, “Supporting the Development of Public Services in Open Government both within and across borders”, Derilinx aims to provide tools and interfaces for intuitive access to open data portals[15] and public services using natural language. Two sub-pilots, consisting of the implementation of a chatbot and an open data dashboard providing an integrated interface, “Kweery”, aim to (1) answer natural language queries regarding public services information via a chatbot and (2) return open data associated with this query via a dashboard. Both the public service information and the open data are returned in the language of the query, providing a web interface for users to access cross-border government services and open data in their native language.

The title of Pilot IV is “Multilingual Text Analytics for Extracting Real-World Evidence in the Pharma Sector”. In this pilot, 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 generating real-world evidence for customers from the pharmaceutical industry. Real-world evidence refers to information on the effectiveness and safety of a drug product that is gathered outside the controlled settings of clinical trials, in order to demonstrate value-add of a drug in terms of improvements in quality of life for specific patient populations. Extracting real-world evidence requires the analysis of large volumes of heterogeneous content, including subjective assessments of patients and medical experts, which is typically available as unstructured text in multiple languages.

1.2.Business Model Canvas (BMC)

Each of the industrial partners was asked to document their pilot project by means of a Business Model Canvas. The Business Model Canvas[16] is a strategic management template for developing new or documenting existing business models. It is a visual chart with elements describing a firm's or product's value proposition, infrastructure, customers, and finances. It assists firms in aligning their activities by illustrating potential trade-offs.

The Business Model Canvas was selected as a documentation tool as it breaks down the business model into easily-understood segments, which are presented in a straightforward, structured way on a single page. This can clarify thinking on the business model. The canvas can be used at any stage of a business so will be useful throughout the life of the Prêt-à-LLOD project.

[15] Initially https://data.gov.ie/ but will be expanded to include the European sites eg. https://datos.gob.es/, the Spanish data portal

[16] Osterwalder, Alexander, Yves Pigneur, Tim Clark, and Alan Smith. Business Model Generation: A Handbook for Visionaries, Game Changers, and Challengers., 2010. Print

Figure 2 depicts the elements of BMC. The canvas has “front” and “back” stages, with the front stage (on the right hand side) showing what drives value and how to reach customers and make money. The back stage (on the left hand side) shows what is required to make the front stage possible. At the centre of the canvas is the value proposition, that is what is being delivered to the customer.

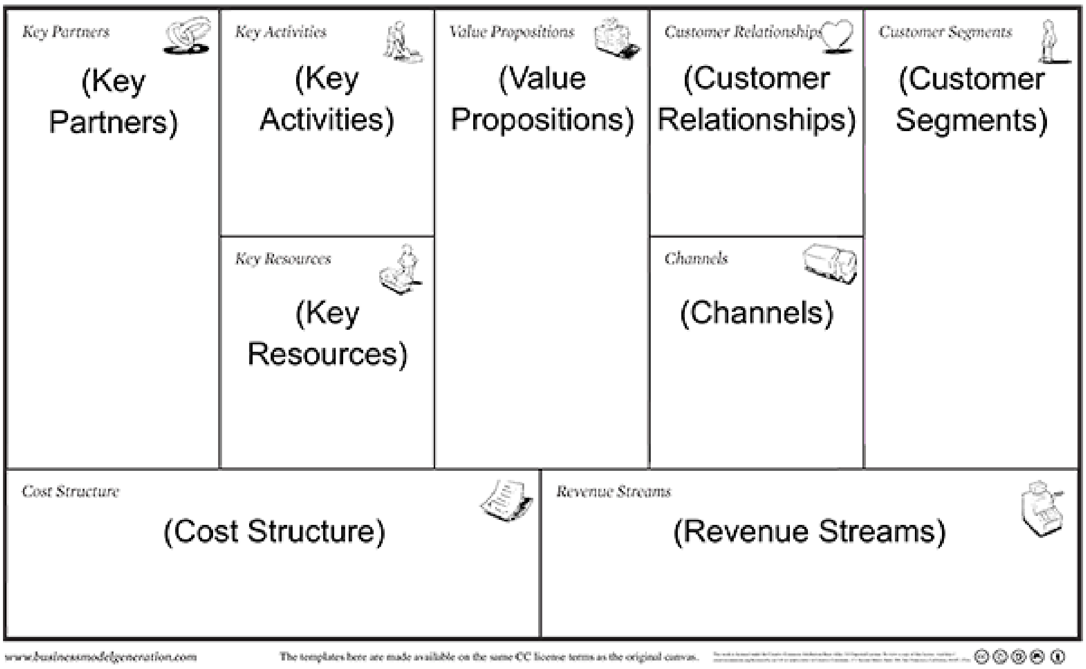


Figure 2: The nine elements of the Business Model Canvas

Customer Segments;

Customer Relationships, the relationships established with the customer;

Channels, the channels through which to reach the customer;

Revenue Streams, revenues generated;

Value Propositions, which is the value of the products or services offered for each segment;

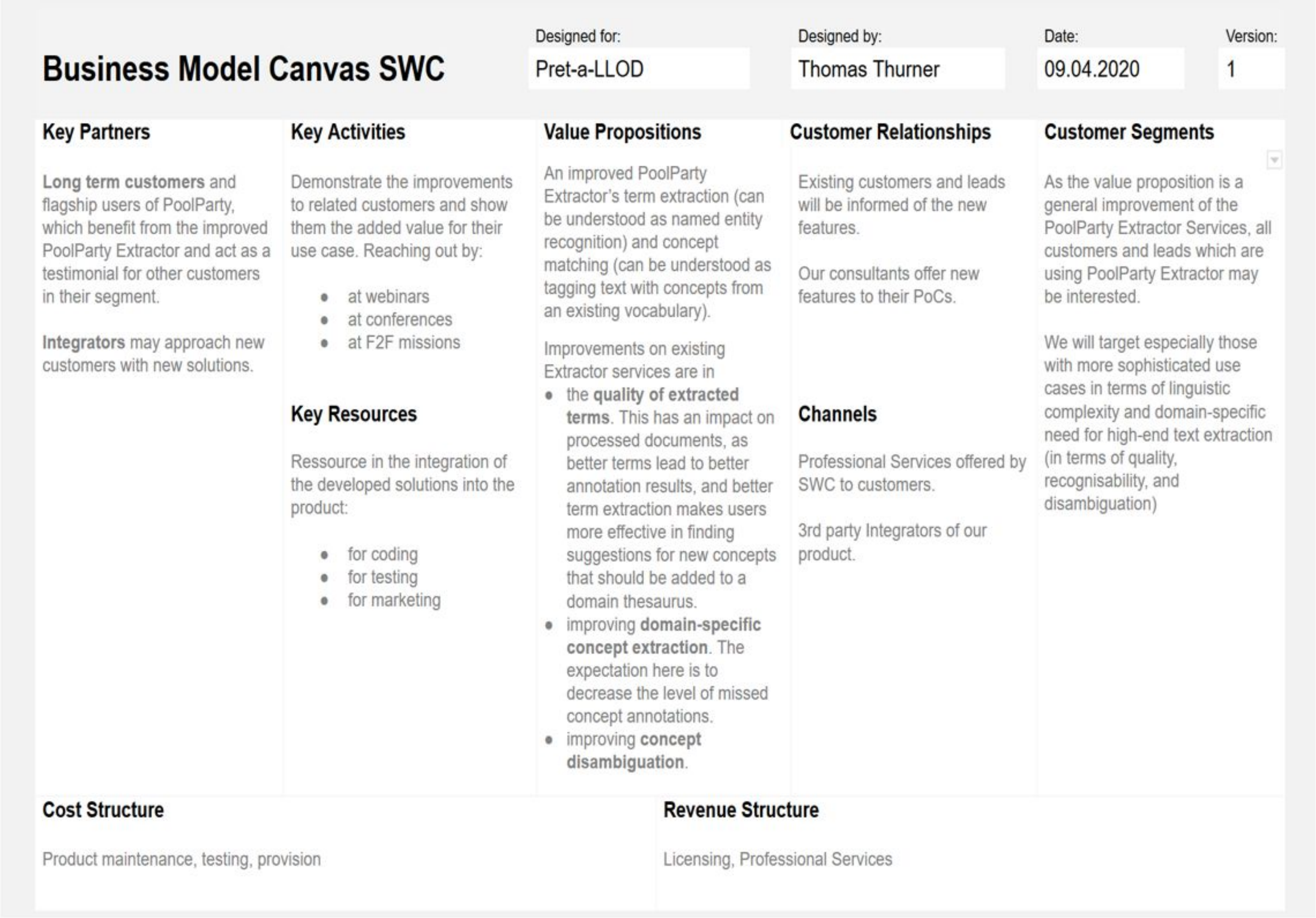
Key Activities, the key activities to make the business model effective;

Key Resources, the organisation/company's key resources;

Key Partners, the key partners with which the company intends to join in order to create value for the customer;

Cost Structure, cost structure for resources, activities and key partners.

2.2.1. BMC for pilot I - Multilingual Knowledge Graphs for Knowledge Management across Sectors



Business Model Canvas SWC

Designed for:	Designed by:	Date:	Version:
Pret-a-LLOD	Thomas Thurner	09.04.2020	1

Key Partners

Long term customers and flagship users of PoolParty, which benefit from the improved PoolParty Extractor and act as a testimonial for other customers in their segment.

Integrators may approach new customers with new solutions.

Key Activities

Demonstrate the improvements to related customers and show them the added value for their use case. Reaching out by:

- at webinars
- at conferences
- at F2F missions

Key Resources

Ressource in the integration of the developed solutions into the product:

- for coding
- for testing
- for marketing

Value Propositions

An improved PoolParty Extractor's term extraction (can be understood as named entity recognition) and concept matching (can be understood as tagging text with concepts from an existing vocabulary).

Improvements on existing Extractor services are in

- the **quality of extracted terms**. This has an impact on processed documents, as better terms lead to better annotation results, and better term extraction makes users more effective in finding suggestions for new concepts that should be added to a domain thesaurus.
- improving **domain-specific concept extraction**. The expectation here is to decrease the level of missed concept annotations.
- improving **concept disambiguation**.

Customer Relationships

Existing customers and leads will be informed of the new features.

Our consultants offer new features to their PoCs.

Channels

Professional Services offered by SWC to customers.

3rd party Integrators of our product.

Customer Segments

As the value proposition is a general improvement of the PoolParty Extractor Services, all customers and leads which are using PoolParty Extractor may be interested.

We will target especially those with more sophisticated use cases in terms of linguistic complexity and domain-specific need for high-end text extraction (in terms of quality, recognisability, and disambiguation)

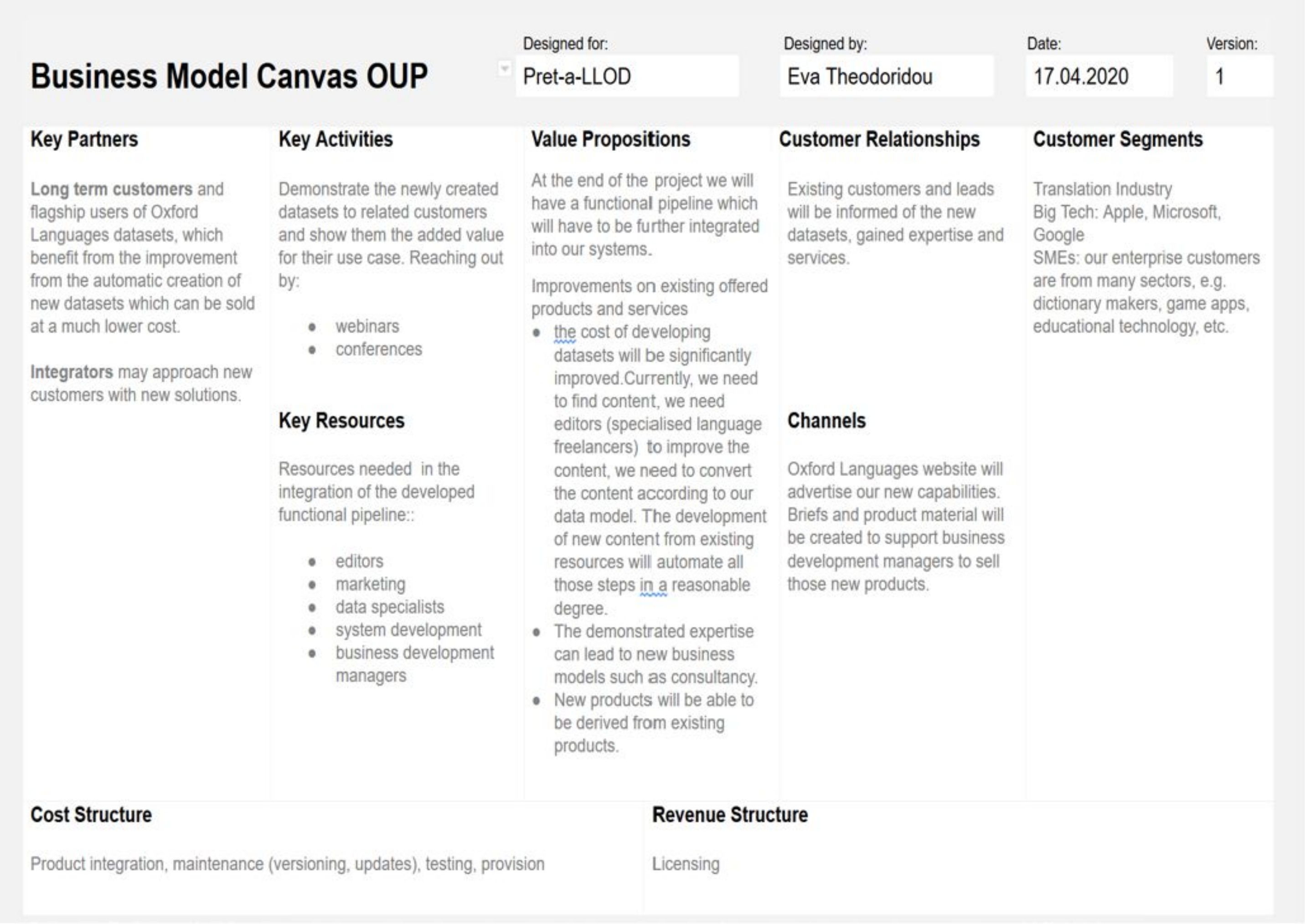
Cost Structure

Product maintenance, testing, provision

Revenue Structure

Licensing, Professional Services

2.2.2. BMC for pilot II - Linking Lexical Knowledge to Facilitate Rapid Integration and Wider Application of Lexicographic Resources for Technology Companies



Business Model Canvas OUP

Designed for: Pret-a-LLOD

Designed by: Eva Theodoridou

Date: 17.04.2020

Version: 1

Key Partners

Long term customers and flagship users of Oxford Languages datasets, which benefit from the improvement from the automatic creation of new datasets which can be sold at a much lower cost.

Integrators may approach new customers with new solutions.

Key Activities

Demonstrate the newly created datasets to related customers and show them the added value for their use case. Reaching out by:

- webinars
- conferences

Key Resources

Resources needed in the integration of the developed functional pipeline::

- editors
- marketing
- data specialists
- system development
- business development managers

Value Propositions

At the end of the project we will have a functional pipeline which will have to be further integrated into our systems.

Improvements on existing offered products and services

- the cost of developing datasets will be significantly improved.Currently, we need to find content, we need editors (specialised language freelancers) to improve the content, we need to convert the content according to our data model. The development of new content from existing resources will automate all those steps in a reasonable degree.
- The demonstrated expertise can lead to new business models such as consultancy.
- New products will be able to be derived from existing products.

Customer Relationships

Existing customers and leads will be informed of the new datasets, gained expertise and services.

Channels

Oxford Languages website will advertise our new capabilities. Briefs and product material will be created to support business development managers to sell those new products.

Customer Segments

Translation Industry
Big Tech: Apple, Microsoft, Google
SMEs: our enterprise customers are from many sectors, e.g. dictionary makers, game apps, educational technology, etc.

Cost Structure

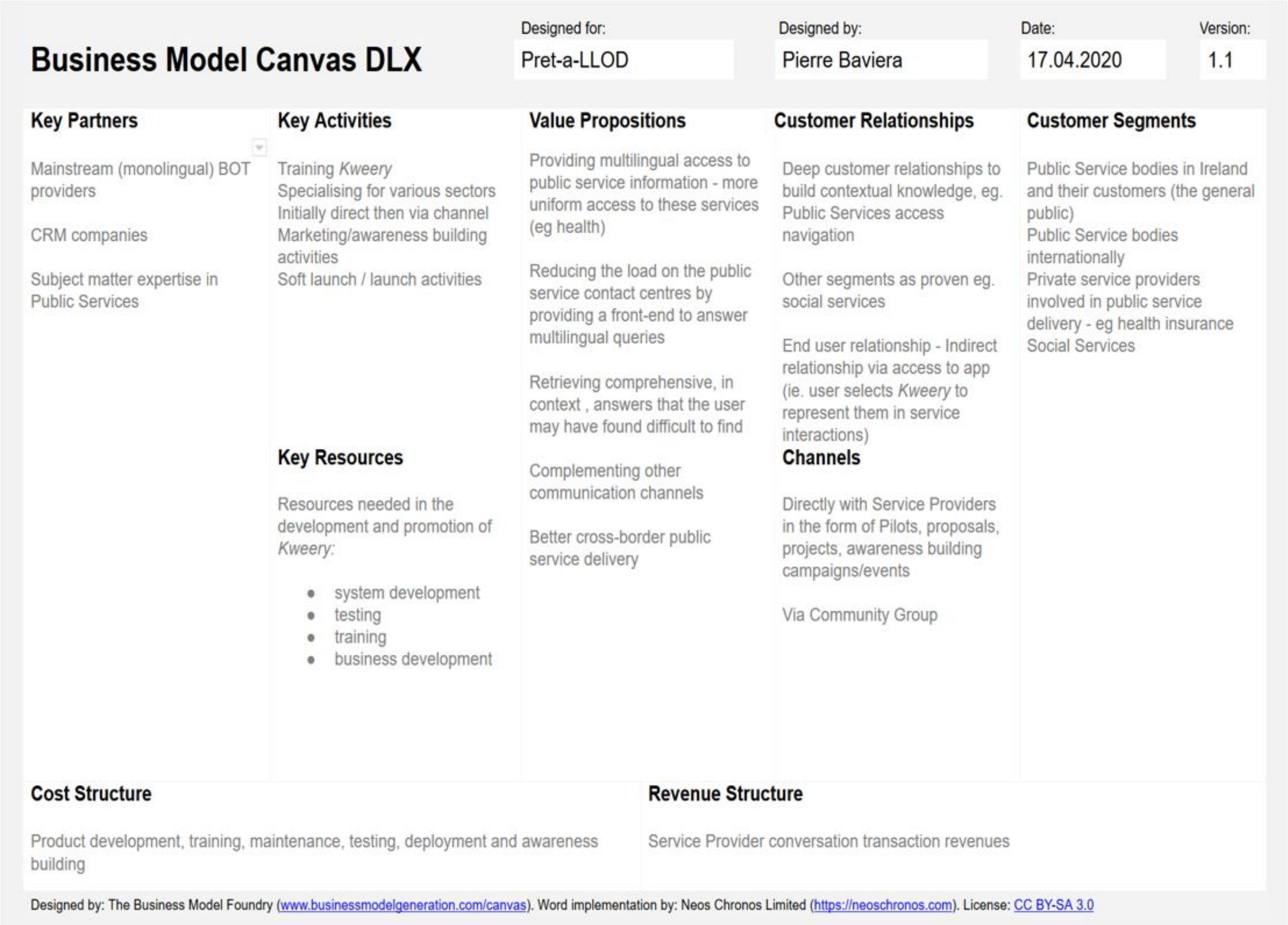
Product integration, maintenance (versioning, updates), testing, provision

Revenue Structure

Licensing

2.2.3. BMC for pilot III - Supporting the Development of Public Services in Open Government both within and across borders

Note: This BMC is documented further in Appendix 1



Business Model Canvas DLX

Designed for: Pret-a-LLOD

Designed by: Pierre Baviera

Date: 17.04.2020

Version: 1.1

Key Partners

Mainstream (monolingual) BOT providers

CRM companies

Subject matter expertise in Public Services

Key Activities

Training *Kweery*
Specialising for various sectors
Initially direct then via channel
Marketing/awareness building activities
Soft launch / launch activities

Key Resources

Resources needed in the development and promotion of *Kweery:*

- system development
- testing
- training
- business development

Value Propositions

Providing multilingual access to public service information - more uniform access to these services (eg health)

Reducing the load on the public service contact centres by providing a front-end to answer multilingual queries

Retrieving comprehensive, in context , answers that the user may have found difficult to find

Complementing other communication channels

Better cross-border public service delivery

Customer Relationships

Deep customer relationships to build contextual knowledge, eg. Public Services access navigation

Other segments as proven eg. social services

End user relationship - Indirect relationship via access to app (ie. user selects *Kweery* to represent them in service interactions)

Channels

Directly with Service Providers in the form of Pilots, proposals, projects, awareness building campaigns/events

Via Community Group

Customer Segments

Public Service bodies in Ireland and their customers (the general public)
Public Service bodies internationally
Private service providers involved in public service delivery - eg health insurance
Social Services

Cost Structure

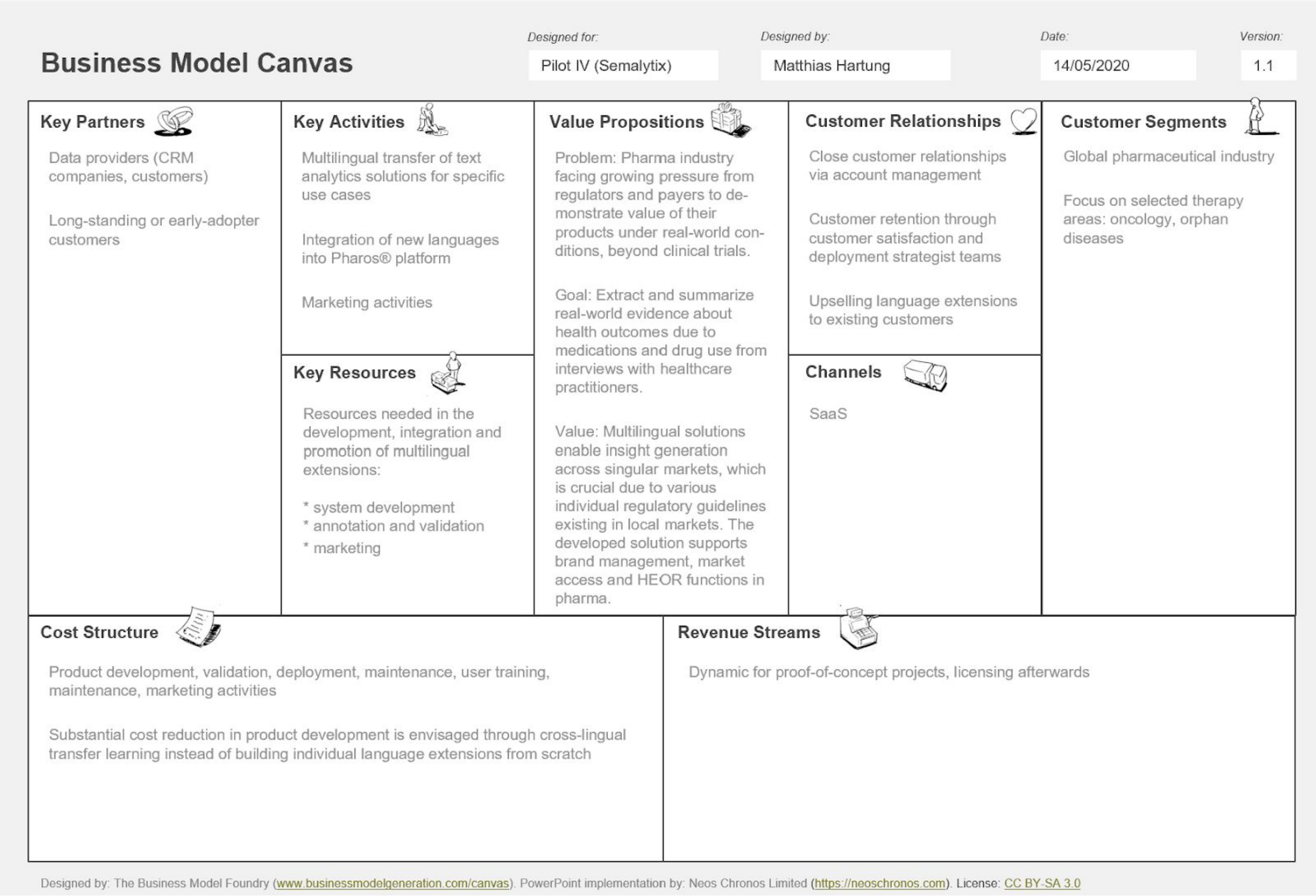
Product development, training, maintenance, testing, deployment and awareness building

Revenue Structure

Service Provider conversation transaction revenues

Designed by: The Business Model Foundry (www.businessmodelgeneration.com/canvas). Word implementation by: Neos Chronos Limited (https://neoschronos.com). License: CC BY-SA 3.0

2.2.4. BMC for pilot IV - Multilingual Text Analytics for Extracting Real-World Evidence in the Pharma Sector



Business Model Canvas

Designed for: Pilot IV (Semalytix)

Designed by: Matthias Hartung

Date: 14/05/2020

Version: 1.1

Key Partners

Data providers (CRM companies, customers)

Long-standing or early-adopter customers

Key Activities

Multilingual transfer of text analytics solutions for specific use cases

Integration of new languages into Pharos® platform

Marketing activities

Key Resources

Resources needed in the development, integration and promotion of multilingual extensions:

* system development
* annotation and validation
* marketing

Value Propositions

Problem: Pharma industry facing growing pressure from regulators and payers to demonstrate value of their products under real-world conditions, beyond clinical trials.

Goal: Extract and summarize real-world evidence about health outcomes due to medications and drug use from interviews with healthcare practitioners.

Value: Multilingual solutions enable insight generation across singular markets, which is crucial due to various individual regulatory guidelines existing in local markets. The developed solution supports brand management, market access and HEOR functions in pharma.

Customer Relationships

Close customer relationships via account management

Customer retention through customer satisfaction and deployment strategist teams

Upselling language extensions to existing customers

Channels

SaaS

Customer Segments

Global pharmaceutical industry

Focus on selected therapy areas: oncology, orphan diseases

Cost Structure

Product development, validation, deployment, maintenance, user training, maintenance, marketing activities

Substantial cost reduction in product development is envisaged through cross-lingual transfer learning instead of building individual language extensions from scratch

Revenue Streams

Dynamic for proof-of-concept projects, licensing afterwards

3. Competitor and Market Analysis

3.1 General market analysis

In any market analysis in 2020, the impact of COVID-19 cannot be ignored. The IMF forecast in April 2020[17] that the global economy is projected to contract sharply by 3 percent in 2020, the impact will be felt most acutely in the advanced economies (see figure 3) with their economies only returning to Q1 2019 levels at the end of 2021.

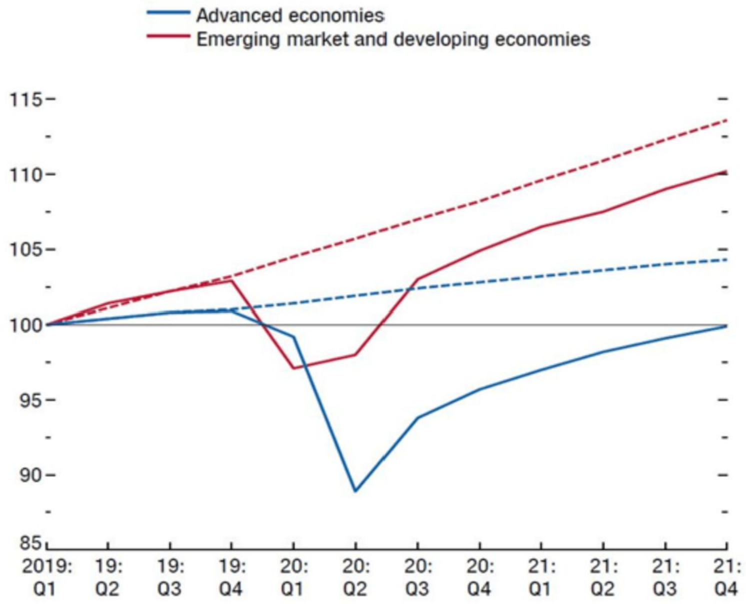


Figure 3: Quarterly World GDP, IMF World Economic Outlook (Dashed lines indicate estimates from January 2020)

In more specific implications for this project, the cancellation of technology conferences, or their moving online, is leading to missed partnership opportunities as in-person networking is no longer possible. In particular, many B2B companies regard in-person events and trade shows as more effective than digital alternatives such as email, product demos or the company website[18] as a channel for sales conversions.

In the translation industry, several post-Corona scenarios are being considered[19]. These range from a contraction of the industry by 30-40% to the opportunities offered by more work being done online offering opportunities for greater innovation and automation.

[17] https://www.imf.org/en/Publications/WEO
[18] https://www.emarketer.com/content/coronavirus-event-cancellation-b2b-marketers
[19] https://blog.taus.net/que-sera-in-the-summer-of-2021

To speed up responses to calls and questions about the virus, NLP and AI in the form of chatbots or digital assistants are being used[20]. A variety of AI-based diagnosis and monitoring tools are playing also a role[21]. This and the fact that most medical and other interactions are now online may lead to increased acceptance levels for NLP- or AI-based applications.

3.2 Potential for growth in AI/NLP services

According to an Accenture survey[22] of European government organisations in Finland, France, Germany, Norway and the UK, the vast majority (86%) of respondents said that their organisation plans to increase its spending on AI in 2020. Most respondents believe that their organisation's leadership is supportive of AI projects, with only one-fifth (21%) reporting a lack of support from the top for such initiatives. The greatest anticipated benefits from these AI investments are increased efficiencies, cost or time savings, and enhanced productivity.

Similarly, Gartner's report into Natural Language Processing Adoption Growth Insights[23] reflected on the adoption of NLP expanding beyond North America, and adoption in EMEA (and specifically Western Europe) increasing by 15 percentage points from 2017-2018.

According to a MarketsandMarkets report of NLP in Healthcare and the Life Sciences [24], market size is projected to grow from USD 1.5 billion in 2020 to USD 3.7 billion by 2025. The ability to analyse and extract meaning from narrative texts and non-related data sources is predicted to be a major driver of this growth.

[20] https://www.techrepublic.com/article/tech-companies-pack-the-frontlines-to-combat-the-coronavirus/

[21] https://www.forbes.com/sites/gilpress/2020/04/17/ai-gets-into-the-fight-with-covid-19/

[22] https://newsroom.accenture.com/news/european-government-organizations-are-enthusiastic-about-artificial-intelligence-but-face-challenges-adopting-it-according-to-accenture-study.htm

[23] Natural Language Processing Adoption Growth Insights, 2019

[24] https://www.reportlinker.com/p04006885/Natural-Language-Processing-Market-by-Type-Technologies-by-Deployment-Type-Vertical-by-Region-Global-Forecast-to.html?utm_source=PRN

3.3. Competitor Overview: Multilingual Knowledge Graphs for Knowledge Management across Sectors (Pilot I) - Semantic Web Company (SWC)

The Enterprise Metadata Management Market was valued at $3.42 billion in 2019. On the basis of the expectation that 90% of new data generated will be unstructured, the market is forecast to reach $11.74 billion by 2025[25]. The market has shown significant growth during 2019, and further such growth can be expected in 2020.

When considering a Metadata Management Solution, a significant portion (more than 20%) of those who opted for SWC's PoolParty had considered IBM, Informatica or Smartlogic[26].

In a general comparison of PoolParty with these 3 competitors, the SWC product received higher general ratings and a higher willingness to recommend.

In the specific categories, PoolParty received higher ratings generally for its Customer Experience than IBM and Informatica and similar ratings to Smartlogic. There are also differences in customer segments, with PoolParty's competitors receiving a very small share of reviews from the Government, Public Service and Education segment (<7%), whereas SWC's solution received 31% of their reviews from this sector.

Semantic Frameworks: In comparing the ratings for their Semantic Frameworks, the IBM and Smartlogic products have a clear advantage, suggesting that this is an area where PoolParty can improve its offering to increase market share. In the Government, Public Sector and Education sector, MS Azure Data Catalog and Oracle Enterprise Metadata Management are key competitors. Both of these receive similar ratings to PoolParty in this category, which suggests that improving this functionality could provide a competitive advantage to SWC.

A recent Gartner report on "Critical Capabilities for Metadata Management Solutions"[27] reports a recent shift in the metadata management market, away from

[25] https://www.researchandmarkets.com/reports/4602282/enterprise-metadata-management-market-size#rela1-4804968

[26] https://www.gartner.com/reviews/market/metadata-management-solutions/vendor/semantic-web-company/product/poolparty/alternatives

[27] https://www.gartner.com/en/documents/3980298/critical-capabilities-for-metadata-management-solutions

focusing on data catalogues to adding more advanced metadata functions. As a result, they see differentiation narrowing between the revenue leaders who offer little more than a data inventory to other products and providers with additional functionality.

The broadening scope of data sources (using metadata from platforms, tools, third-party providers and a widely divergent range of data sources and user experiences) along with the introduction of “active metadata” concepts means that the basic functionalities no longer differentiate solutions.

Gartner’s “Critical Capabilities for Metadata Management Solutions” report states, “Simply referred to as “active metadata management,” this fast-growing approach to metadata has emerged as the key to utilizing all data in any organizational, regulatory or ecosystem solution. Active metadata management includes information for determining the context and semantic interpretation of data. Finally, it has the potential for dynamic resource allocation and system optimization.”

In Gartner’s ranking of SWC against its 16 “Magic Quadrant”[28] competitors, SWC is ranked third for their approach to “active metadata management”.

3.4. Market Trends: Multilingual Knowledge Graphs for Knowledge Management across Sectors (Pilot I) - Semantic Web Company (SWC)

Market Trends according to Gartner 2020 [28]

Due to market- and competitor-sensitive information, “Magic Quadrant for Metadata Management Solutions” from Gartner is quoted from in this section:

> Although the market opportunity can appear extremely large — after all, metadata is everywhere — the success of metadata management as a distinct discipline for delivering value to organizations is not yet secure. Moreover, success with metadata management must be supported by technological evolution and, more importantly, by changes in metadata practice. It also requires the participation of a wider set of roles, including business roles, in the metadata management process.

[28] Gartner Magic Quadrant for Metadata Management Solutions, 16 October 2019 G00372820, Analyst(s): Guido De Simoni | Mark Beyer | Ankush Jain, https://www.gartner.com/document/3970385

To fulfill the market’s demand for easy management of data, despite an increasingly complex data landscape, metadata management solution vendors must do more than describe and provide transparency regarding the usage of data. They must also become active players in managing data. Vendors are adjusting to and exploiting the following changes:

- The transfer of metadata ownership from the CIO to the chief data officer (CDO) or a similar role
- The increase in the variety and extent of metadata supported
- The enhancement of the scope of metadata through automation (ML) and through automated enrichment by semantic search capabilities, standard processes and crowdsourcing
- The rise of semantics formalism (also known as formal ontologies) for improved interoperability
- The development of shared understanding across multiple domains
- New ways to capture and visualize metadata (driven by data preparation for analytics)

We have already seen, during the past 12 months, a focus on the pervasive use of metadata. This focus relates specifically to the pervasive use of metadata (business and technical, but also statistical and audit-related) in data management technologies ranging from database to data integration and even data quality technologies. The result is increasing automation of many activities, such as database optimization and tuning, data integration and data preparation, and detection and implementation of rules for data governance. All these activities will make extensive use of descriptive metadata, which has been the focus so far, and will turn it into active metadata (see Note 1), which, in turn, will lead to the automation of many data management implementation and maintenance activities. Metadata management vendors have a unique opportunity to play a key role in collecting, analyzing and sharing metadata from the overall data management landscape. They could then turn this metadata into actions — either by directly implementing these activities or, more realistically, by sharing the insights generated by active metadata with partners such as DBMS [database management system] vendors, data integration vendors, data quality vendors and even MDM [master data management] vendors.

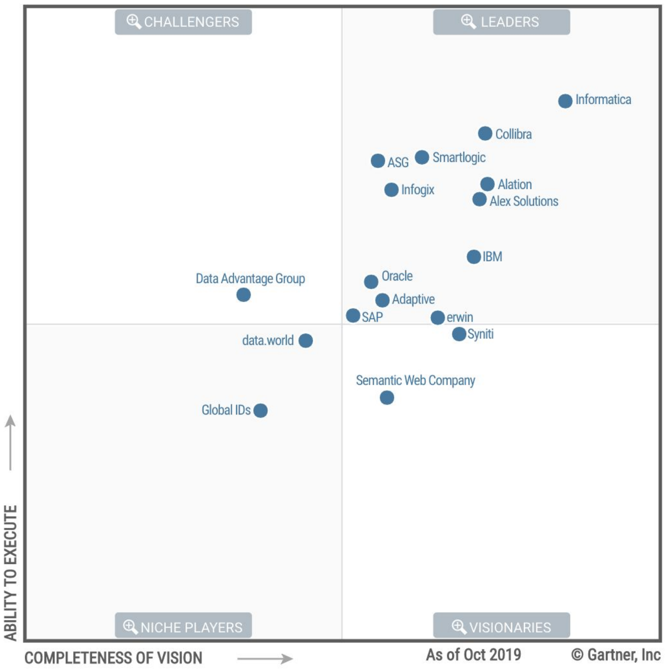


Figure 4: Magic Quadrant for Metadata Management Solutions

Overall, it is important to note that there are still some inhibitors of faster adoption of metadata management solutions. They include:

- The lack of maturity of strategic business conversations about metadata (see “Create a Business Case for Metadata Management to Best Fulfill Your Data and Analytics Initiatives”).
- The fact that metadata management is still a nascent discipline in most organizations and accounts for only 12% of the time spent on data management (see “The State of Metadata Management”).
- The expensive but required effort to integrate metadata management solutions in multivendor environments. This inhibitor has started to be addressed, however, by new vendors’ initiatives relating to openness and interoperability (see, for example, ODPi).
- The lack of identification of accurate metadata management solutions whose capabilities meet the current and future requirements of specific use cases.

Most organizations will find that their current metadata management practices differ across applications, data and technologies, and that these practices are siloed by the needs of different disciplines — each with their own governance authority, practices and capabilities. Data and analytics leaders who have already invested in data management solutions should first evaluate the capabilities of their existing solutions — including federation/integration capabilities, support for ML and AI, and cloud options — before buying a new solution. However, if they are dealing with emerging use cases — including collaborative analytics and community-oriented data and analytics governance — they must also assess new metadata management solutions, driven by active metadata, that are fuelling the convergence of other data management disciplines.

3.5. Competitor Overview: Linking Lexical Knowledge to Facilitate Rapid Integration and Wider Application of Lexicographic Resources for Technology Companies (Pilot II) - Oxford University Press (OUP)

The end product of the OUP Pilot is an innovative way to generate bilingual dictionaries. There are many companies and institutions that offer bilingual dictionaries. Traditional dictionary production is primarily manual, with content being created by human lexicographers and translators. There are also companies that use more automatic means (like machine translation) to generate new content, however the content that we have evaluated so far (e.g. lexicala) is of significantly lower quality.

Hence, as competitor analysis we should focus more on the technological advancements (means of creating dictionaries) rather than competitors.

Technological advancements

The situation in the **lexical linking area**, and particularly in what concerns the use case of interest here, namely the ability of automatically generating new dictionary content, still poses some challenges. The most notable recent activity on this front is that developed within the *Translation Inference Across Dictionaries* (TIAD) framework. The shared task there consists in “exploring methods and techniques for automatically generating new bilingual (and multilingual) dictionaries from existing ones”, which aligns very well with the use case for **OUP Pilot**, although the approach taken there differs from ours: while TIAD task considers only linking across bilingual dictionaries, we take a monolingual dictionary as the hub where all bilinguals link to, and which provides a sort of central inventory of senses.

There is a new on-going TIAD edition this year (TIAD 2020) for which there are no results available yet. However, last year's edition (Gracia at al. 2019) showed that the area is still far from a satisfactory level of maturity. Most remarkably, none of the participating systems were able to improve the baselines determined by the organizers, one of which was in fact from seminal work in the area carried out in the 90s (Tanaka & Umemura 1994). Table 2 provides the best results for each participating team.

	Precision	**Recall**	**F1**
Baseline 1 (Tanaka et al. 1994)	0.64	0.26	0.37
Baseline 2 (word2vec)	0.66	0.24	0.35
Frankfurt	0.64	0.22	0.32
LyS-DT	0.36	0.31	0.32
UNLP-NMT-3PATH	0.66	0.13	0.21
ONETA-ES	0.81	0.10	0.17

Table 2: Results for the baselines (highlighted in blue) and the best system submitted by each participating team.

At present within the field of lexical linking, there is also a further line of work going on as part of the EU-funded project ELEXIS (*European Lexical InfraStructure*). Of particular relevance here is the globalLex 2020 track developed within that framework, *Monolingual Word Sense Alignment Shared Task*, which is still in progress and for which there are no results available yet. It requires a system to identify whether 2 definitions from the same lexeme in different dictionaries for the same language correspond to the same sense.

There are two elements that make this task very comparable to the work carried out by OUP in Pilot II. Firstly, the fact that it involves sense relations between dictionaries of the very same language, as opposed to the TIAD framework, where the sense alignment takes place across languages. Secondly, the possible types of relation between the two definitions, which can be: "exact", "broader", "narrower", "related to" or "none". While the latter ("none") corresponds to the "non-link" class returned by the OUP basic sense linking system (in opposition to the "link" class), the former 4 are equivalent to the "perfect", "wider than", "narrower than", and "partial" distinctions determined by Pilot II sense granularity classifier.

In spite of these points of connection, however, there is a significant difference between this shared task and the work in OUP Pilot II. Namely, the fact that the shared task targets sense linking between two monolingual dictionaries, whereas OUP sense linking system focuses on aligning a monolingual and a bilingual dictionary. The difference is not trivial because bilingual dictionaries do not feature definitions, which is the piece of information used in the shared task for identifying whether the sense alignment between dictionaries applies. This element makes it difficult to directly take advantage of any significant development from the shared task without re-engineering the components developed so far at OUP to some extent, i.e., the basic sense linking tool, its complementing quality estimator, and the sense granularity classifier.

3.6. Market Analysis: Linking Lexical Knowledge to Facilitate Rapid Integration and Wider Application of Lexicographic Resources for Technology Companies (Pilot II)

Market segments

For the last couple of months, OUP have been investigating the following market segments:

- Language services;
    - Machine translation (prioritised)
    - Language education
    - Language localisation
        - Gaming
        - Software localisation
- CX platforms:
    - Chatbots
    - Text classifiers
- Other types of software: text editors, ML engines and NLP services, content management software, question answering software, summarisation software

Market problems

The investigation of the above market segments has led to the identification of several market problems in regards to language (training) data; which are listed below. Not all of them are applicable for every market segment but they are widely shared.

- Lack of resources for languages other than: English and main European
- Lack of resources for domain specific data (e.g. specialised terms)
- Bias in data: gender, regional, political, etc.
- Quality of the available material, which are noisy, include duplicates, do not have any type of linguistic or semantic tagging.
    - Training models has a huge carbon footprint: so quality is vital for the environment since we can end-up with less iterations if models are trained with higher quality of data
- Lack of expertise dealing with language data
- Lack of enough volume (1M tokens) to train MT/NLP algorithms
- Available data within organisations is locked up in legacy formats rendering them unusable in the modern scenarios of machine translation

3.7. Competitor Overview: Supporting the Development of Public Services in Open Government both within and across borders (Pilot III) - Derilinx

The Derilinx pilot will provide access to public service information via a dashboard and a chatbot. In addition, the dashboard will present related open data.

In both the US and Europe, dashboards are used to present progress on Open Data policy. Dashboards can be a key feature of smart city performance where they present information on the city environment, and can be used to compare cities across a country[29]. Cityflow[30] is a commercial enterprise providing integrated devices and visualisations. Envisio[31] supports local government in developing public dashboards to monitor progress against objectives.

In general, however, there is not significant usage of dashboards for presenting open data so this competitor analysis focuses on the chatbot functionality.

The main providers of chatbot frameworks are IBM Watson, Google Dialogflow, Amazon Lex and Microsoft Azure. These provide similar features at similar prices:

[29] https://smart-cities.dashboard.gov.au/all-cities/overview

[30] https://www.cityflow.live/

[31] https://envisio.com/solutions/strategic-plan-dashboard/

	NLP	Multilingual	Pricing
Amazon Lex	NLU to extract intent	Amazon Translate service (support for 54 languages)	$0.00075 per text request $15.00 per million characters (translation)
Dialogflow (Google)	Google tools	supports 20 languages (with variants)	$0.004 per request
IBM WatsonAssistant	AI, ML, NLP and reasoning	Watson language translator supports 38 languages	Standard package $0.0025 per message
Microsoft Bot Framework	Microsoft LUIS (Language Understanding Intelligent Service)	LUIS supports 17 languages; alternative is MS translator API	$0.50 per 1,000 messages (premium channels)

Although all of the above provide NLP and ML tools, support for multiple languages and translation support, none of them can provide the direct multi-lingual (cross-border) search capability with the option of including linked open data:

- Question posed in original language
- Internally, question is translated to target language and information retrieved from cross-border site
- The response is presented to the user in their original language
- Linked open data is included in the response

Additionally, the API-based design of the Derilinx pilot allows for new Prêt-à-LLOD functionalities to be incorporated as they become available.

3.8. Market Analysis: Supporting the Development of Public Services in Open Government both within and across borders (Pilot III)

Multilingual access to Government and Health Information services

In an analysis of access to European government services, 78% provided multilingual access to their government services portal through language selection at the portal level. This ensures the quality of the information across the secondary/alternative languages, in contrast to where tools such as Google Translate are provided for the user.

Every site reviewed that has multilingual access provides English as an alternative to their primary language(s). This means that access to these services for English speakers is generally very good, but can lead to the exclusion of those without English. In particular, in the UK and Ireland, services are provided through English (and Welsh in Wales and through Irish in Ireland).

State websites in the US providing access to state and health services were also reviewed, with very few (Texas, California) providing official information in Spanish and English. Of the rest, most (64%) rely on Google translate (but with an official disclaimer in some cases). In the case of New Mexico health services, they have added multilingual access in Spanish and Vietnamese to their English-only website, this is for COVID information only.

Chatbot access to Government services

According to Slator[32], European consumers are more receptive to chatbots than those in the US; 50% of French consumers hold a positive opinion on bots as opposed to only 32% of Americans.

Of the European government service websites analysed, two (Finland and Portugal) provide multilingual chatbot access, both including English as an alternate language alongside their main official languages (Finnish/Swedish and Portuguese respectively). Another example is found in Dubai, where the government provide a multilingual chatbot[33], supporting Arabic and English, to answer customer queries regarding the Dubai Electricity and Water Authority.

[32] https://slator.com/technology/multilingual-chatbots-the-conversation-is-yet-to-get-longer/

[33] https://www.dewa.gov.ae/en/rammas

Chatbots are more prevalent in the US states, with 28% providing this access to their public services. In light of the COVID-19 crisis, Utah offers a chatbot specifically to answer questions regarding the pandemic.

Chatbots and Open Data

The city of Kansas has created a Facebook chatbot[34] with the express purpose of making its open data portal more accessible to non-technical users. There is also a company in Spain, Xatkit, providing chatbots that access all the standard open data formats used by governments (Socrata[35], CKAN[36] and ODATA[37]) as the best way for non-technical people to access and benefit from open data.

Proficiency in English

The latest Irish Census (2016) includes in a summary of the top 10 non-Irish nationalities their own assessment of their ability in English[38]. Those who have assessed their ability as "not well, not at all or haven't stated", come to a total of over 58,000. This is out of 535,475 non-Irish nationals usually resident in Ireland, a significant proportion of whom (103,113) are from the UK and so can be assumed to have ability in English.

Looking at the top 10 non-Irish nationalities alone, this suggests that of non-Irish nationalities excluding those with English as a first language (268,862), there could be some 22% who are significantly disadvantaged by their ability in English and so have limited access to services.

Similar analysis of the figures for England and Wales[39] indicates that a significant number of those with a main language other than English regard themselves as speaking English "not well" or "not at all" - see figure 5.

Main language	Total population aged 3 and over	Non-proficient

[34] https://chatbotslife.com/4-government-agencies-using-chatbots-faf98702b775

[35] https://www.tylertech.com/products/socrata

[36] https://ckan.org/

[37] https://www.odata.org/

[38] https://www.cso.ie/en/releasesandpublications/ep/p-cpnin/cpnin/introduction/

[39] https://www.ons.gov.uk/peoplepopulationandcommunity/culturalidentity/language/articles/detailedanalysisenglishlanguageproficiencyinenglandandwales/2013-08-30

Polish	546,174	150,618
Panjabi	273,231	88,604
Bengali	221,403	67,336
Urdu	268,680	63,231
Gujarati	213,094	50,414
Chinese other than Cantonese and Mandarin	141,052	34,690
Arabic	159,290	28,042
Portuguese	133,453	25,646
Spanish	120,222	12,493
French	147,099	8,332

Figure 5: Top 10 largest populations for main languages other than English by non-proficiency in English

Language Proficiency and Health

The latest census for England and Wales included questions designed specifically to investigate how the self-reported ability to speak English correlates to general health. Around 300,000 residents aged 3 and over in England and Wales who could not speak English well or at all reported their general health as 'Not Good'. Also, people with a main language other than English who could not speak English well or at all had a lower proportion of 'Good' general health (65 per cent) than those with English as their main language (80 percent), or those with a main language other than English who spoke English well or very well (88 per cent).

Conclusion

As demonstrated above, there is significant interest and potential for providing cross-border multilingual access to health information and government services, particularly for some of the less well-served languages. In order to identify the languages which would have maximum impact, local census data can be used. Covid-19 has highlighted the importance of reaching the largest possible audience by providing health and government information in multiple languages.

3.9. Competitor Overview: Multilingual Text Analytics for Generating Real-World Evidence in the Pharmaceutical Domain (Pilot IV) - Semalytix

A) Consulting Companies

Company	Focus	Business Model	Life Science Exclusivity	Use Cases	Data Sources
Cello Health	Market Research Business Insights and Analytics for Pharma	Consulting	Yes	Patient Research: patient journey work, disease burden scoping, dialogue analysis, support service optimisation and ecosystem mapping	Multiple Channels, also Social Media
Kantar Health	Data-driven Consulting Services based on established research frameworks	Consulting based on Hero Framework, BrandPlus Framework, Claritis	Yes	HEOR[40], Commercial Effectiveness, Corporate Reputation, Brand Positioning, Clinical and Scientific Assessment, Late Phase Research, Market Opportunity Assessment	Kantar's Patient-Centered Research (PaCeR) database with clinical data from electronic health records, labs, medical and pharmacy claims

B) Platform Providers

[40] Health Economics and Outcomes Research

Company	Focus	Business Model	Life Science Exclusivity	Use Cases	Data Sources	Multilinguality
Aetion	Real-World Evidence Generation	Platform	Yes	**Life Sciences Companies:** optimizing R&D and advancing strategies for regulatory approval **Payers:** determine effective treatments for specific patient populations and measure impact in outcomes, utilization, and cost	claims, electronic health records, registries, and clinical trial data	No
Palantir	Real-World Evidence Generation	Platform (Palantir Foundry)	No	Discovery, Therapy Cost-Effectiveness, Patient Population Dynamics, Drug Outcomes	pre-clinical, clinical, manufacturing, sales, and marketing data	No

Signals Analytics	AI-driven advanced analytics platform connects all the relevant data sources	BI platform	No	Market Research: Uncover Competitive Strategies, Prioritize R&D Pipeline, Surface Early Innovation, Identify Promising Partners & Assets	patent filings, research papers, conference programs, clinical trials, drug listings and more	No
Quid	Get faster insights to inform strategic decisions.	Platform	No	Voice of the Patient, Key Opinion Leaders, CRM Analytics	patient forum conversations and drug reviews	No
Sensyne Health	Use AI to detect hidden patterns in anonymised patient data, accelerating the development of new medicines	„Docking Station“	Yes	Supporting pharmacologic discovery, clinical trials support and analysis, RWE for medicines on the market	NHS Data	No

C) Technology Providers

Company	Focus	Business Model	Life Science Exclusivity	Use Cases	Data Sources	Multilinguality

Linguamatics **(acquired by IQVIA)**	Several products incl. **Voice of the Customer NLP Services**	NLP Services	Yes	Life Science Companies	Customer Calls	Yes

The cases of **Cello Health** and **Kantar Health** show that there is a demand for data-driven high-quality consulting services for Real World Evidence generation, comprising the needs to inform benefit-risk assessment, support payer negotiations, optimization of product development and treatment and care pathways, as well as understanding patient trajectories. In contrast to Semalytix, they follow a traditional consulting and project-based business model.

Aetion and **Palantir** focus on semi-structured and more traditional data sets such as: claims, electronic health records, registries, and clinical trial data, manufacturing data as well as manufacturing and sales data. These data sets only allow for a very limited understanding of patients' needs and trajectories.

There are also technology vendors such as **Linguamatics** that deliver NLP services and technological solutions.

The closest competitors to Semalytix are: **Quid, Sensyne Health** and **Signals Analytics**. Sensyne and Signals work on complementary data sources. Quid and Semalytix are very comparable in terms of data sources and use cases. Semalytix has a technological competitive edge and analytical superiority allowing to analyse patient trajectories at higher level of depth and granularity. Also, Quid does not support multilingual processing of data sources. **Sensyne Health** has access to NHS patient records and understands itself as a "docking station" matching between NHS data and companies seeking to exploit the data. Signals Analytics has a different focus on data sources such as: patent filings, research papers, conference programs, clinical trials, drug listings, and others.

By considering benefit-risk assessments through the perspective of health care practitioners, Semalytix can capture aspects of the real-world treatment experience that are not captured in clinical contexts, as for instance by Sensyne. In that regard, Semalytix and Sensyne Health have complementary, synergetic offerings.

Semalytix' competitive edge: deliver insights at the level of quality that life science companies are used to from consulting agencies such as Cello Health and Kantar, for non-traditional and unstructured data sources at machine speed and at unrivaled quality and depth compared to other competitors.

3.10. Market Analysis: Multilingual Text Analytics for Generating Real-World Evidence in the Pharmaceutical Domain (Pilot IV)

In the following, we analyse the market potential of multilingual text analytics for generating real-world evidence (RWE) for the pharmaceutical industry along the two dimensions of (i) relevance and market size of real-world evidence in pharma, (ii) regional foci of the global pharmaceutical market and the languages involved.

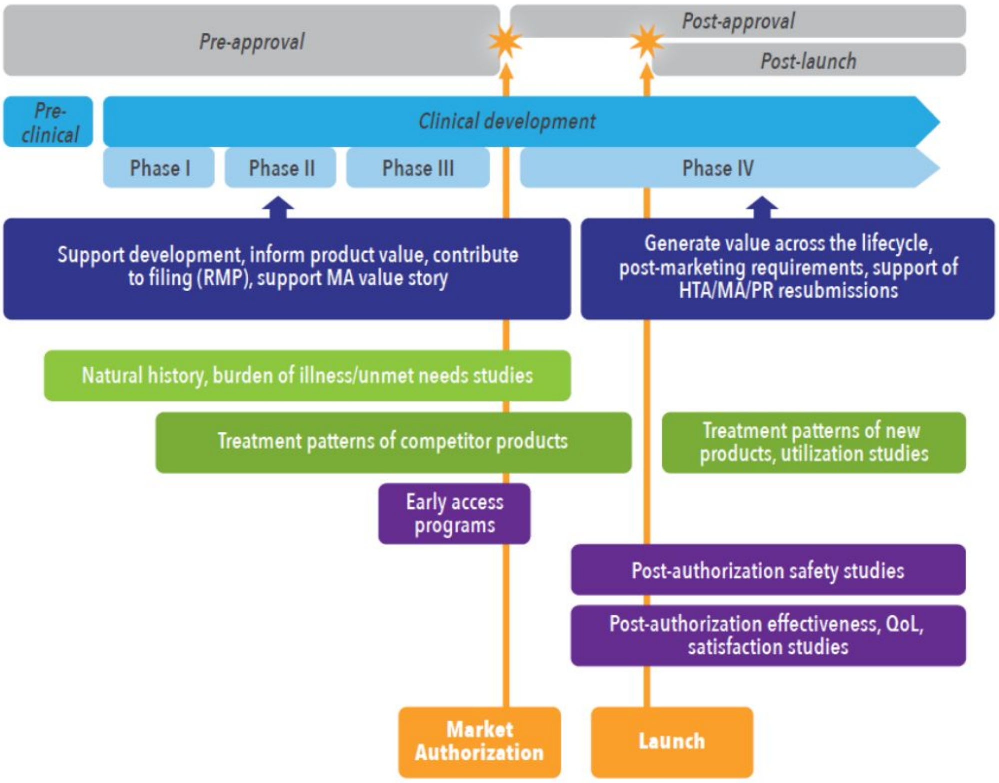
According to the Natural Language Processing Adoption Growth Insights Report published by Gartner in January 2020[41], the interest in NLP is on the rise across many verticals, with generally high potential for NLP technologies in Healthcare, Utilities and Government. In terms of regional uptake, the adoption of NLP is expanding beyond North America. These potentials can be quantified more precisely based on the MarketsandMarkets report from April 2020 quoted above[42], which predicts the global Natural Language Processing (NLP) in healthcare and life sciences market to grow from USD 1.5 billion in 2020 to USD 3.7 billion by 2025, at a Compound Annual Growth Rate of 20.5% during the forecast period. As major growth factors, the report identifies the increasing demand for improving Electronic Health Records (EHRs) data usability to better patient care, and ability to analyse and extract meaning from narrative texts and other unstructured data sources.

The quoted high-impact data sources are precisely the ones which hold the strongest potential for generating real-world evidence, i.e., assessing the value of drugs and medical interventions outside the controlled conditions of clinical trials. In particular, the interest in clinical narratives, either from medical experts' or the patients' perspectives, marks a shift towards an increased use of "non-traditional" data sources for RWE generation (contrary to the "traditional" RWE sources such as electronic health records or claims data). Irrespective of its provenance, RWE is considered to have strong impacts on various stages of the pharmaceutical product lifecycle, as can be seen from the following figure taken from an Evidera white paper [43].

[41] https://www.gartner.com/en/documents/3978977/natural-language-processing-adoption-growth-insights-201

[42] https://www.reportlinker.com/p04006885/Natural-Language-Processing-Market-by-Type-Technologies-by-Deployment-Type-Vertical-by-Region-Global-Forecast-to.html

[43] https://www.evidera.com/protocol-design-in-real-world-evidence-the-indispensable-link-between-strategic-need-and-study-execution/

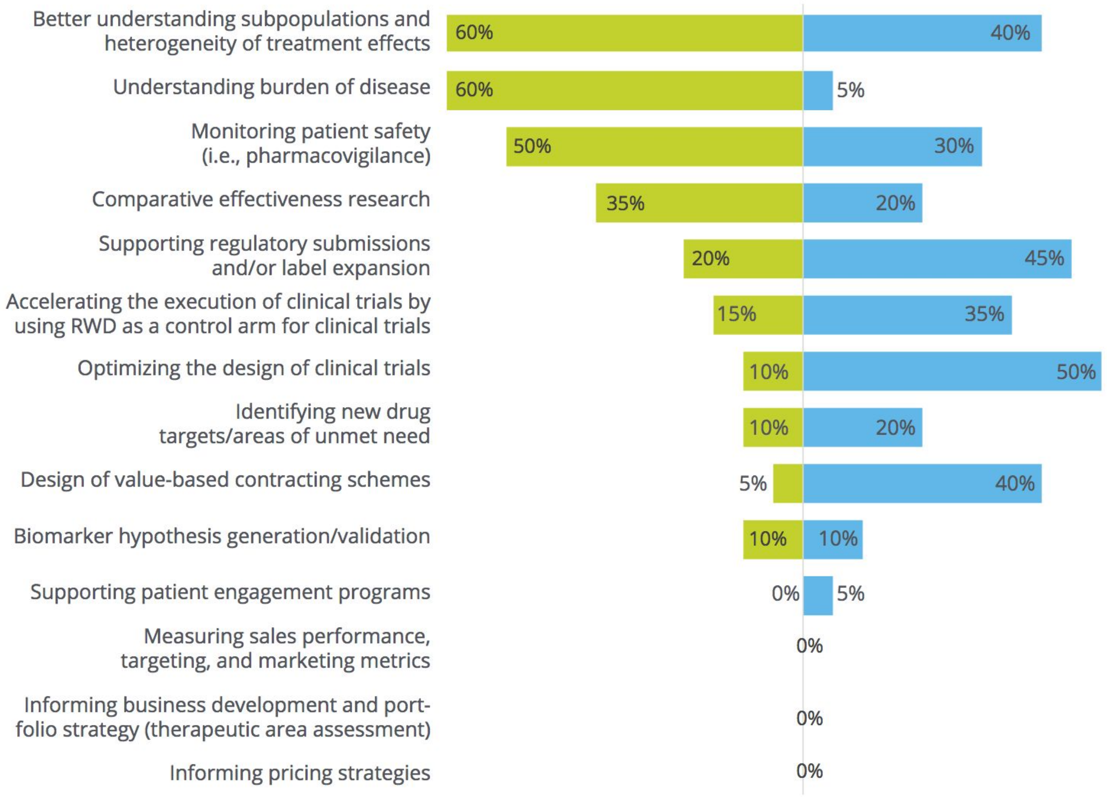

Pre-approval
Post-approval
Post-launch
Pre-clinical
Clinical development
Phase I
Phase II
Phase III
Phase IV
Support development, inform product value, contribute to filing (RMP), support MA value story
Generate value across the lifecycle, post-marketing requirements, support of HTA/MA/PR resubmissions
Natural history, burden of illness/unmet needs studies
Treatment patterns of competitor products
Treatment patterns of new products, utilization studies
Early access programs
Post-authorization safety studies
Post-authorization effectiveness, QoL, satisfaction studies
Market Authorization
Launch

HTA = health technology assessment; MA = market access; PR = pricing and reimbursement; QoL = quality of life; RMP = risk management plan

Figure 6: Objectives of Real-World Evidence across Therapeutic Product Development and Lifecycle

In a survey[44] on applications of RWE that are considered most impactful, Deloitte asked stakeholders from pharmaceutical companies to rank the areas displayed in the figure below with respect to their perceived current (green bars) and future (blue bars) impact. Some of the aspects rated as most relevant (primarily subpopulation analysis, monitoring safety and effectiveness, identifying unmet needs, or optimizing clinical trials by selecting appropriate endpoints) are directly addressed as part of the value proposition offered by Semalytix Pharos®.

[44] https://www2.deloitte.com/content/dam/insights/us/articles/4354_Real-World-Evidence/DI_Real-World-Evidence.pdf



Note: The figure denotes current and future application areas ranked amongst the top three by respondents and expressed as a percentage.

Source: Deloitte's 2018 RWE Benchmarking Survey.

Deloitte Insights | deloitte.com/insights

Figure 7: Perceived current (green) to future (blue) impact of RWE

The continuously increasing importance of RWE in drug development cycles is also underlined by regulatory uptake. Based on the 21st Century Cures Act from 2016, the Food and Drug Administration (FDA) is charged with "evaluating the expanded use of RWE, including its potential to support the approval of new indications for previously approved drugs".[45] In the meantime, the FDA has issued several guidances for the pharmaceutical industry on how to use real-world evidence to support regulatory decision-making for medical interventions and devices.[46] As a consequence, a number of regulatory examples have recently occurred in which RWE has been utilised to support regulatory decisions either at authorization or to support an extension of indication.[47] Based on these cases and further evidence, Olson (2019)

[45] https://www2.deloitte.com/us/en/pages/life-sciences-and-health-care/articles/real-world-evidence-benchmarking-survey.html

[46] https://www.fda.gov/science-research/science-and-research-special-topics/real-world-evidence

[47] Cave, Alison et al. (2019): Real-World Data for Regulatory Decision Making. Challenges and Possible Solutions for Europe. Clinical Pharmacology and Therapeutics 106(1): 36-39. https://doi.org/10.1002/cpt.1426

concludes that the economic potential of RWE amounts to potential savings of 1 billion USD for the global pharmaceutical industry per year, provided that “evidence-powered operating frameworks” can be established in pharma organisations in order to ensure coherent and integrated evidence generation.[48]

As a rough approximation of the market volume that could be targeted by solutions offering real-world evidence generated from textual sources, about 20% of the pharma respondents in the Deloitte survey say that they are or will be using social media as data source in order to elicit non-traditional RWE (for which the FDA has also issued guidance materials under the patient-focused drug development program [49]).

The need for multilingual approaches in generating RWE from non-traditional textual sources arises from the fact that the global pharmaceutical market is fragmented into multiple singular regional markets with individual regulatory guidelines existing and different languages spoken in each of those. According to recent market statistics[50], the total revenue of the world-wide pharmaceutical market in 2019 amounts to 1033 billion USD. While the largest share (47.5%) is in the US market, European countries and Japan are the largest markets involving languages other than English (accounting for more than 20% of the global revenue together). The rest of the global revenue is generated in emerging markets including countries like China, Russia, Brazil and India. In particular, the highest individual growth rates are exhibited by the Chinese pharmaceutical sector over the previous years. In all these emerging markets as well, native languages other than English are spoken, which means that multilingual approaches are required in order to generate RWE from non-traditional sources that are tailored to the needs of these markets.

Conclusion

In the above analysis, we have demonstrated the relevance and economic potential of generating RWE which is quantified as an annual savings potential of 1 billion USD within the global pharmaceutical industry, provided that integrated evidence generation frameworks are established. In such frameworks, non-traditional textual sources for RWE generation will have their role to play, and given that approximately 40% of the annual revenue (400 billion USD) are currently generated in non-English

[48] Olson, Melvin (2020): Can real-world evidence save pharma US$ 1 billion per year? A framework for an integrated evidence generation strategy. Journal of Comparative Effectiveness Research 9(2): 79-82. https://www.futuremedicine.com/doi/pdf/10.2217/cer-2019-0162

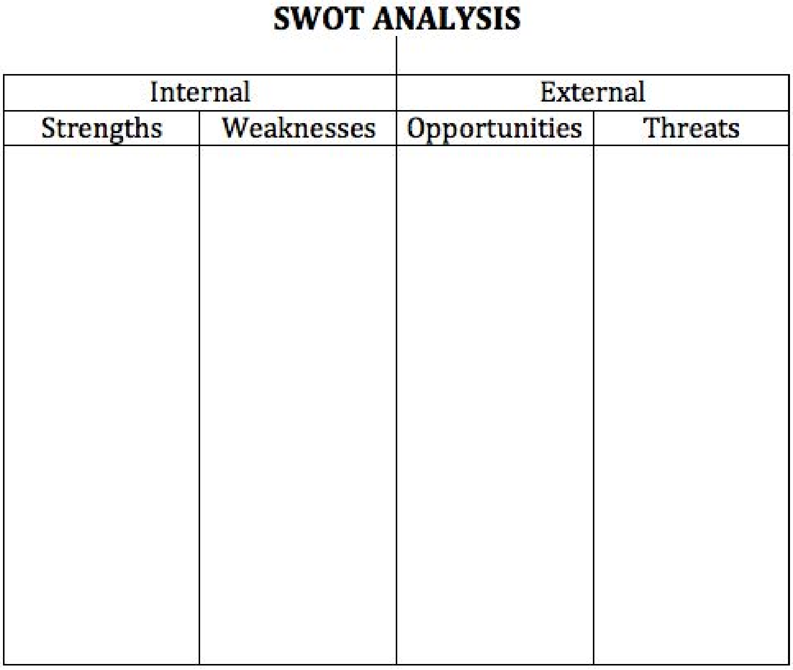
[49] https://www.fda.gov/drugs/development-approval-process-drugs/fda-patient-focused-drug-development-guidance-series-enhancing-incorporation-patients-voice-medical

[50] https://www.statista.com/topics/1764/global-pharmaceutical-industry/

speaking countries, this clearly demonstrates the strong need and business potential for multilingual LT solutions in this area.

4. SWOT analysis

The SWOT analysis is a strategic planning tool used to evaluate Strengths, Weaknesses, Opportunities and Threats of a project or in a business or any other situation where an organisation has to make a decision to achieve a goal. SWOT analysis assesses internal and external factors, as well as current and future potential.



SWOT ANALYSIS

Internal		External	
Strengths	Weaknesses	Opportunities	Threats

Figure 8: Swot Analysis Model

The analysis of strengths and weaknesses (see figure 8) is used to determine distinctive internal competencies that will distinguish the organisation from the rest of the market. The weaknesses could also be used as opportunities for adjustment of the project.

Looking outside, to the market, the analysis of opportunities and threats should reveal aspects that the project can use to its advantage to improve its competitive position.

The analysis of the context of the Pret-a-LLOD project has revealed some strengths and weaknesses of the project. At the same time, potential threats and opportunities from external factors and scenarios have emerged.

The SWOT Analysis summary table is presented below:

INTERNAL FACTORS	
STRENGTHS	WEAKNESSES
● Open source software, open standards ● Academic partners' experience ● Industrial partners' participation in a range of commercially important fields ● Project's focus on addressing the significant problem of a lack of standardization (across both industry and academia) ● Closer collaboration between academia and industry ● Pilot projects providing opportunities to improve industrial partners' product offerings ● Several pilots involve domain or task adaptation workflows which often make a crucial	● Inflexibility of project budget ● (Potential) project extension causing issues with industry-academic collaborations

difference in practical usability for business use cases, thus increasing the likelihood of customer uptake (compared to offering rather “generic” LRs[51] or services)	
EXTERNAL FACTORS	
OPPORTUNITIES	**THREATS**
● European Commission Digital Single Market strategy ● Market advancements: Government and Healthcare are among the sectors displaying most interest in NLP [52] ● Technological advancements: Advances in NLP have increased potential for linking lexical sense data with corpora[53]	● Similar product/services offered by competitors ● Impact of COVID-19 on purchasing budgets ● Market trends that could potentially lead to open source initiatives: collaboration in times of crisis[54], collaboration due to market demands (NLP, MT carbon footprint[55]) ● Potential entry barriers: companies may have their own workflows in place already, may enable some market problems to be resolved through alternative means

[51] Language Resources

[52] https://www.gartner.com/en/documents/3978977/natural-language-processing-adoption-growth-insights-201

[53] Breit, A. A. Revenko, K. Rezaee, M. T. Pilehvar, Jose Camacho-Collados (submitted) WiC-TSV: A Multi-Domain Benchmark for Disambiguating Words in Context. To appear.

[54] https://md.taus.net/corona : community is collaborating in linking corpus data and freely offering it.

[55] https://www.technologyreview.com/2019/06/06/239031/training-a-single-ai-model-can-emit-as-much-carbon-as-five-cars-in-their-lifetimes/: the increased impact that NLP and ML has to the environment can motivate people towards sharing data

5. Conclusion

This initial Business Development Report has given the industrial partners an opportunity to evaluate and elaborate on the business strategy for their pilots and also to learn from other partners' plans. Since the project is at an early stage, each partner will be involved in learning from their initial results and adjusting their plans to achieve the optimum sustainable business model. Market changes will also have a significant impact and the SWOT and Competitor Analyses will be updated accordingly. As the technologies and standards to be delivered by other work packages in the Prêt-à-LLOD project become available, they will be incorporated into the pilots. As the pilots are developed the partners will validate and refine each component of their Business Model Canvas.

Market changes and any changes to the Business Models will be reflected in the subsequent deliverable D2.4, Strategic Report on Business Development v2 (M30) where the final business models will be elaborated. D2.4 will also document the incorporation of the deliverables from the other Prêt-à-LLOD work packages into the pilot products.

Appendix

A1. Detail on Pilot II Business Model Canvas

Further detail was developed on this business model and is documented here.

Kweery - Supporting the development of cross-border public services in open government

1. Customer Segments

Customers (paying for the service):

The Targeted Customers for the Derilinx Pilot are Public Service organisations. These organisations will benefit most from end-users accessing the Kweery multilingual chatbot effectively reducing the volume of complex call/correspondence with Customer Service Front-Line Departments and leveraging analysis of the interaction data to understand better the needs of Public Service users.

The end-users of the service will include citizens (international) that are in the process of gathering information linked to accessing Public Services from another country (e.g. in the context of a move), policy makers in Health for instance (in the context of International benchmarking or addressing a particular individual case arriving/leaving the country), private sector organisations including Health Insurers (on-boarding new customers, advising customers on international cover or setting-up a new policy scheme). End-users will also include those living in a country where they do not have a strong grasp of either English or their host country's language.

Languages to be covered by the Pilot include:

- English
- Spanish

Further to the completion of the Pilot, the service will be extended to additional Public Services including Revenue, Social Welfare, Education, Tourism, Housing/Local Authorities etc. Selection of additional languages will be done further to additional market research.

2. Customer Relationships

Co-creation relationship in the initial phase: Development of specific sector capabilities to train the specialised Kweery chatbot will require deep partnerships with a number of potential customers or Subject Matter Experts.

Post this initial training phase, this will move to a self-service type of relationship.

The relationship with the end user will be indirect as they will access the service through a Client app.

3. Channels

Product awareness will be developed via a dedicated online page completed by referral calls to decision makers in the relevant Public Service Organisations (participation in relevant community forums and/or support via economic development agencies will also be leveraged). Public Service Organisations will support advertising of the service.

A demo version of the pilot will be made available online for evaluation purposes.

Individual Public Service providers will purchase the service directly via an initial training/set-up fee followed by an annual subscription fee.

The service will be delivered online via a dedicated app.

A support service will be included as part of the annual subscription fee with a dedicated online support chatbot.

4. Revenue Structure

Revenue will be secured in two ways:

- Set-up / training fees for the initial stage of set-up for each new customer
- Annual subscription fee to cover access to the service and maintenance/support activities

5. Value Proposition

The volume of cross-countries/cross jurisdiction moves has increased significantly over the past 50 years.

According to the UN, about 272 million people, or 3.5% of the world population, were living outside their country of birth in 2019[56], with Europe hosting the largest number of international migrants (82 million).

[56] https://www.un.org/development/desa/en/news/population/international-migrant-stock-2019.html

Whilst a majority of Government Services web sites include a local language and usually some information in English (see Market Research), citizens that do not master either of these languages struggle to gather and assimilate important Public Service Information.

The Covid-19 crisis has highlighted the disconnect between the language of communication linked to Public Service Providers and citizens/immigrants not mastering the local language.

The volume of information requests across multiple languages (representing complex communication) and jurisdictions has significantly increased and puts pressure on Public Service Contact Centres (phone, emails) as traditional single language – single jurisdiction chatbots cannot handle these types of queries.

The Value Proposition that Kweery offers is a multilingual chatbot capability to address this challenge by:

- Handling queries/build responses in a selection of European languages
- Handling queries regarding multiple European destinations
- Retrieving data from the European Data Portal and other open data portals within Europe
- Being based on an API so can be easily deployed to any platform
- Delivering results via an integrated chatbot and dashboard

Benefits to Public Service Providers (will be expanded to cover Health Service Providers post pilot):

- Provide high availability of Public Service information
- Multilingual support
- Analysis of questions provides data to identify:
    - unmet needs
    - most popular questions
- Opportunity to enrich the response with:
    - open data
    - similar topics that may be of interest

Benefits to the users/citizens include:

- Finding information easily and quickly

- Getting responses in their own language
- Time savings

6. Key Activities

Key activities to complete the Pilot project include:

- Initial development of Kweery chatbot and dashboard general functionalities
- Sector specific training of chatbot, leveraging historical information from the Public Services (Ireland) for instance. Includes testing.
- Awareness/Marketing campaign across a selection of Public Service Organisations in order to demonstrate benefits
- Effective engagement with Spanish Public Service for further chatbot training an
- Soft launch activities with group of users across both languages/jurisdictions

7. Key Resources

Key resources to deliver the pilot include:

- System Development resources including UX (Web App, Chatbot), API, Core Functionality (NLP, Information Extraction, Query Builder etc), API Development, Data Modelling
- Public Services sector Subject Matter Expert to assist in building relationships with potential customers as well as assisting in the training and testing of Kweery
- Selected groups of potential users across English/Spanish language for testing
- Representatives of Public Services in Ireland and Spain to provide access to training data

8. Key Partners

Key partners will include:

- For the System Development phase, a review of monolingual Bot to assess their performance and potential re-usability for Pilot purposes
- For the pilot, Public Service Organisations for training and testing of Kweery

9. Cost Structure

Main cost drivers for the pilot include:

- Application Development
- Testing
- Maintenance
- Initial Project Set-Up with new Public Service Provider
- Customer Support Costs
- Marketing / Business Development / Account Management Costs