D2.4 - Strategic Report on Business Plan Development v2

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

Date: 30th June 2021

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

D2.4 - Strategic Report on Business Plan Development v2

Deliverable Number: D15
Dissemination Level: P(ublic)
Delivery Date: M30
Version:
Author(s): Pierre Baviera (DLX), Katharine Cooney (DLX), Matthias Hartung (SEM), Katherine Martin (OUP), Thomas Thurner (SWC), Thierry Declerck (DFKI)

Document History

Version Date	**Changes**	**Authors**
5/5/2021	First draft	DLX
15/6/2021	Updated sections	DLX
18/6/2021	Updated sections	OUP
21/6/2021	Updated conclusion	DLX
24/6/2021	Internal Review	DFKI

1 Executive Summary

This deliverable presents the second version of the strategic business development plan for Prêt-à-LLOD, developing and reporting the commercial partners' specific business plans and achievements from version 1 of the report, delivered in month 18.

The Prêt-à-LLOD commercial partners are:

Semantic Web Company[1] (SWC)

Oxford University Press[2] (OUP)

Derilinx[3] (DLX)

Semalytix[4] (SEM)

All Prêt-à-LLOD commercial 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 pilot projects of the Prêt-à-LLOD project.

Updates to version 1 of the deliverable have been made in the following sections:

1.3.1 Impact on Technology companies

2.3 Business Model Canvas for pilot II – Linking Lexical Knowledge to Facilitate Rapid Integration and Wider Application of Lexicographic Resources for Technology Companies

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

3.1 General market analysis

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

3.3.4 Market Analysis: Pilot II - Oxford University Press (OUP)

3.3.5 Competitor Overview: Pilot III - Derilinx

3.3.6 Market Analysis: Pilot III - Derilinx

[1] https://semantic-web.com/
[2] https://global.oup.com/
[3] https://derilinx.com/
[4] https://www.semalytix.com/

4.1 SWOT Analysis for Prêt-à-LLOD

5 Collaboration with Academic Partners

6 Conclusion

Contents

Executive Summary 2

1. Introduction 6

1.1. Background 6

1.2. Relationships with other project deliverables 8

1.3. Expected impacts 8

1.3.1. Impact on Technology companies 9

1.3.2. Impact on Pharma 9

1.3.3. Impact on Health Information and Government Services 10

2 Commercial partners' pilots 10

2.1 Business Model Canvas (BMC) 11

2.2 BMC for pilot I - Multilingual Knowledge Graphs for Knowledge Management across Sectors 14

2.3 BMC for pilot II - Linking Lexical Knowledge to Facilitate Rapid Integration and Wider Application of Lexicographic Resources for Technology Companies 15

2.4 BMC for pilot III - Supporting the Development of Public Services in Open Government both within and across borders 16

2.5 BMC for pilot IV - Multilingual Text Analytics for Extracting Real-World Evidence in the Pharma Sector 17

3 Competitor and Market Analysis 18

3.1 General market analysis 18

3.2 Potential for growth in AI/NLP services 19

3.3 Competitor and Market Analysis 19

3.3.1 Competitor Overview: Pilot I - Multilingual Knowledge Graphs for Knowledge Management across Sectors - Semantic Web Company (SWC) 19

3.3.2 Market Trends: Pilot I - Semantic Web Company (SWC) 21

3.3.3 Competitor Overview: Pilot II - Linking Lexical Knowledge to Facilitate Rapid Integration and Wider Application of Lexicographic Resources for Technology Companies - Oxford University Press (OUP) 23

3.3.4 Market Analysis: Pilot II - Oxford University Press (OUP) 25

3.3.5 Competitor Overview: Pilot III - Supporting the Development of Public Services in Open Government both within and across borders - Derilinx 26

3.3.6 Market Analysis: Pilot III - Derilinx 29

3.3.7 Competitor Overview: Pilot IV - Multilingual Text Analytics for Generating Real-World Evidence in the Pharmaceutical Domain - Semalytix 33
3.3.8 Market Analysis: Pilot IV - Semalytix 37
4 SWOT analysis 42
4.1 SWOT Analysis for Prêt-à-LLOD 43
5 Collaboration with academic partners 45
6 Conclusion 46
7 Appendix – References 47

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 more concrete details on the revenue generation model for each pilot utilizing the business canvas method and presents a picture of the current situation and conclusions as to next steps.

1.1. Background

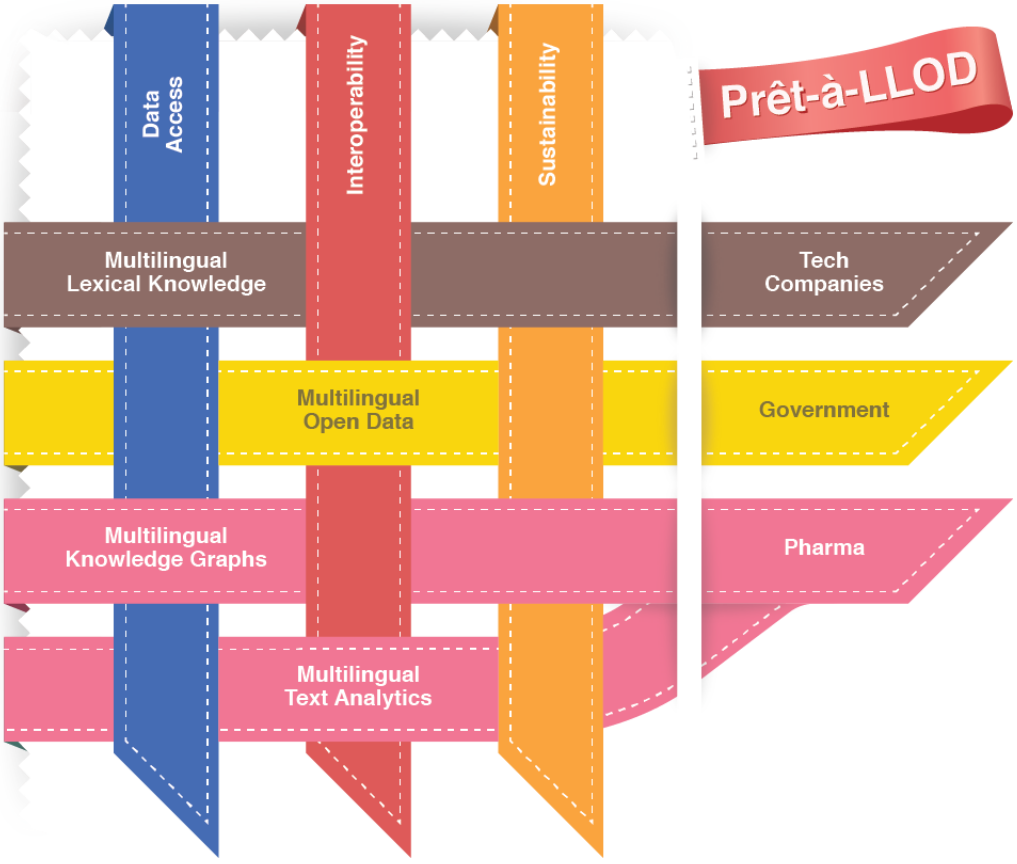


Figure 1: Overview of objectives, uses cases and sectors for Prêt-à-LLOD

This diagram depicts the interaction between the commercial 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 commercial 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 commercial 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 was 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 was submitted in M26 of the project, reported on the progress of 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.

- 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

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

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 "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 commercial partners in reducing costs and time-to-market for their products.

1.3.1. Impact on Technology companies

Software companies rely on high quality lexical resources for a variety of applications, including display to end users, support of functionality in products, and as an input for natural language processing. As they expand their offerings beyond English, the relevant lexical resources become more difficult to identify, source and prepare. Making lexical 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 lexical 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

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

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.

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 are focusing 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 are using their pilot to provide cross-border and multilingual access to health information and government services.

2 Commercial 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 are looking 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

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

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

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

the language technology industry, have used this pilot to implement a novel methodology for generating bilingual dictionaries.

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 information from open data portals 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.

2.1 Business Model Canvas (BMC)

Each of the commercial partners was asked to update the Business Model Canvas they documented in the previous version of this report. 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.

[15] Initially https://data.gov.ie/ but has been modified to return data from the European Open data portal, https://data.europa.eu/en

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

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.

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 generate profits. The backstage (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.

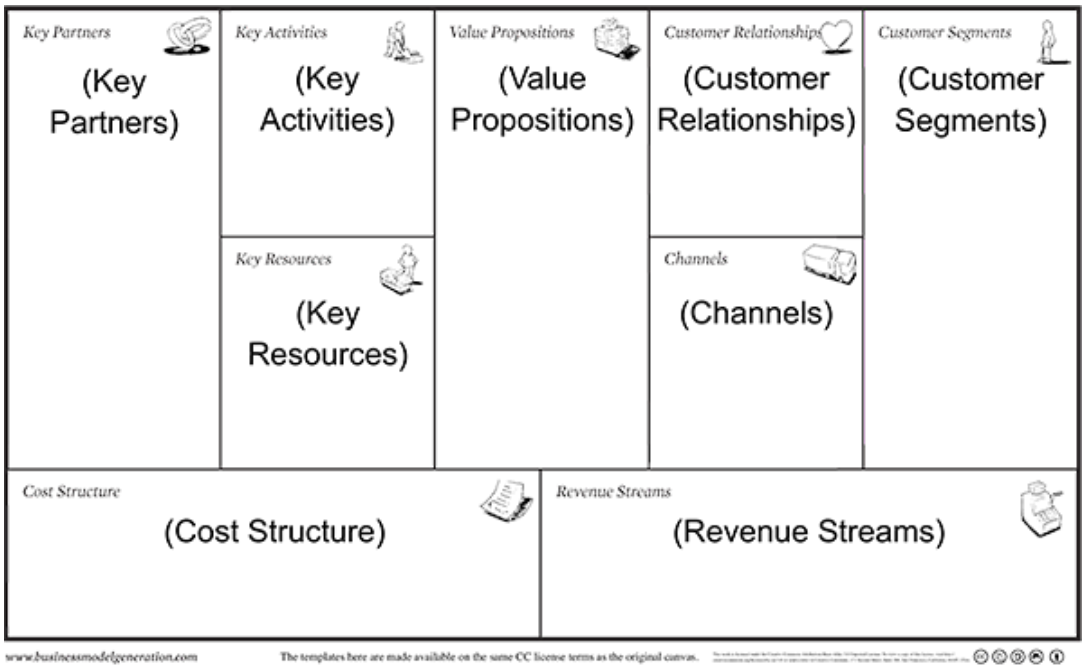


Figure 2: The nine elements of the Business Model Canvas

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 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

2.3 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:	Designed by:	Date:	Version:
	Pret-a-LLOD	Eva Theodoridou (v1)	17.06.2021	2

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:

- lead generation campaigns
- 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 by enabling automatic creation of new language data combinations for post-editing by experts.
- The demonstrated expertise can lead to new business models such as service provision.
- 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.
New language data combinations may enable Oxford Languages to enter new territories and sectors with new products.

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.
Lead generation campaigns can be conducted via social media and advertising

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

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

Business Model Canvas DLX

Designed for:	Designed by:	Date:	Version:
Pret-a-LLOD	Pierre Baviera	18.06.2021	2

Key Partners

CRM companies

Subject matter expertise in Public Services

University of Bielefeld

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

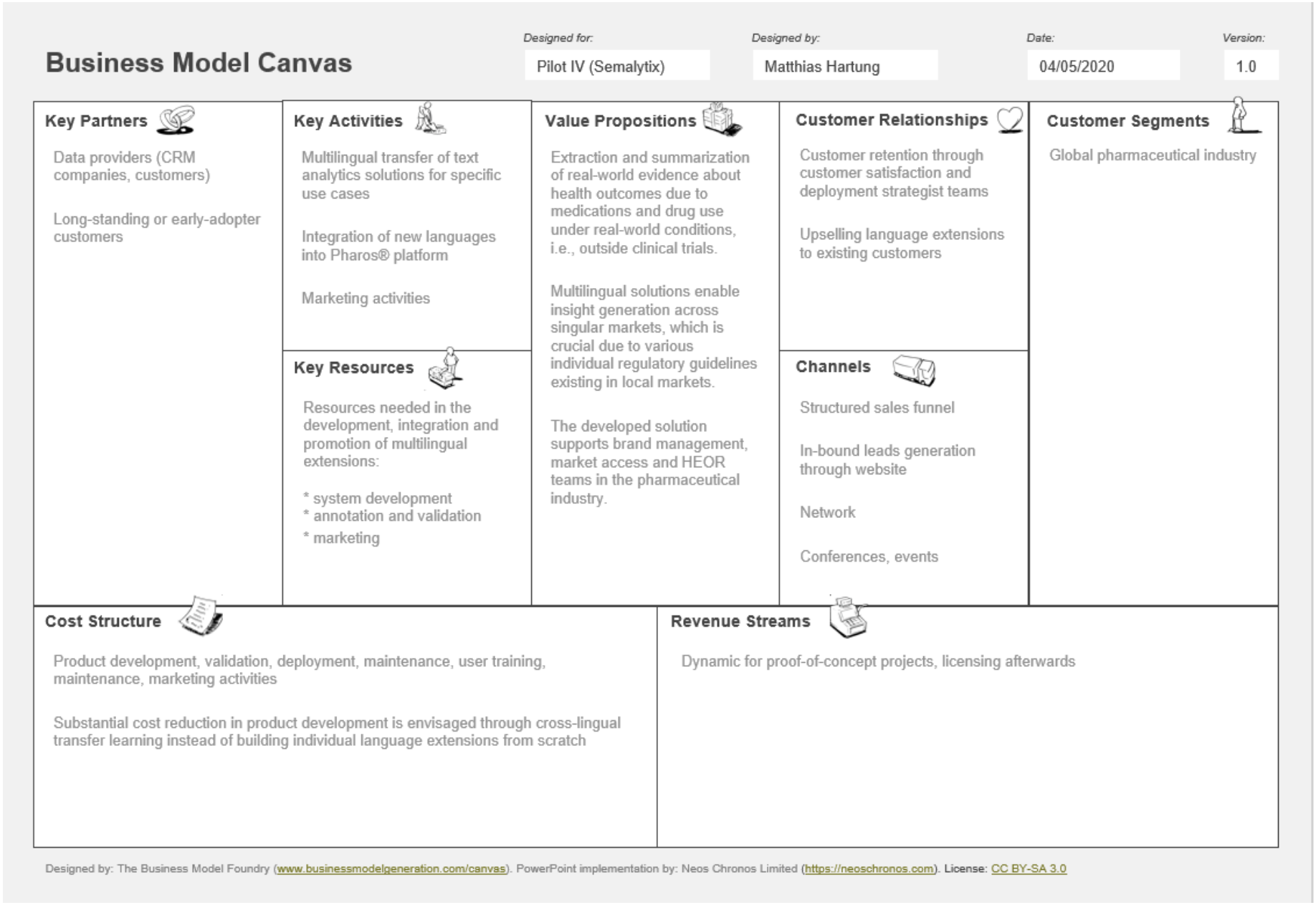
Cost Structure

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

Revenue Structure

Service Provider conversation transaction revenues

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



3 Competitor and Market Analysis

3.1 General market analysis

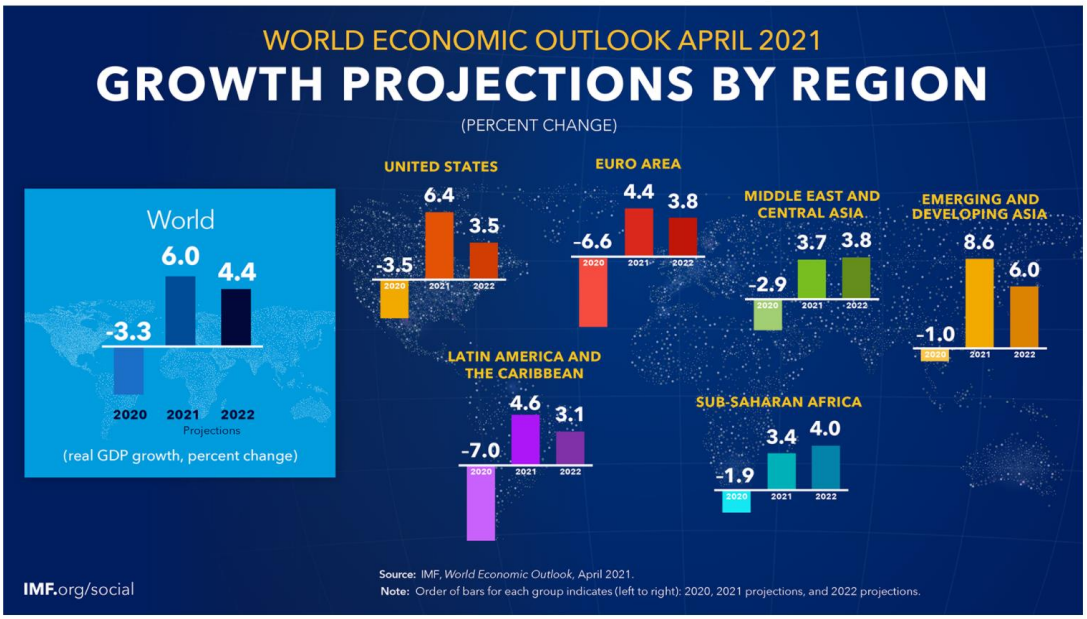


Figure 3: IMF World Economic Outlook (April 2021)

The International Monetary Fund predicts growth in 2021 in the Euro area of 4.4%[17], following a contraction of 6.6% in 2020. The improvement in the projection from that made in October 2020 is due to additional fiscal support in a few large economies, the anticipated vaccine-powered recovery in the second half of 2021, and continued adaptation of economic activity to reduced mobility. This is, of course, highly dependent on the path of the pandemic.

Stanford's AI index[18], an annual study of AI impact and progress developed by a team at Stanford's Institute for Human-centered Artificial Intelligence (HAI) with organizations from industry, academia, and government, reports that despite the pandemic's economic hit, AI investment and hiring increased. "The biggest increases were in healthcare and pharma, where four times as many people increased their investments as decreased them," Erik Brynjolfsson, of the steering committee of the study, says. However, sectors beyond health care, including education, retail, and automotive, also showed increased AI investment.

According to the World Economic Forum[19] (WEF), the COVID-19 pandemic has helped widen global usage for chatbot technology. Many companies and organizations are leading the charge in deploying chatbots to provide COVID-19 information. The two most authoritative voices of the pandemic, the World Health Organisation and Centre for Disease Control, have also included chatbots in their websites to provide up-to-date information to billions on the spread of the disease and its symptoms. Many governments are also launching chatbots to provide validated information to their citizens.

[17] https://www.imf.org/en/Publications/WEO/Issues/2021/03/23/world-economic-outlook-april-2021

[18] AI Index 2021 | Stanford HAI

[19] https://www.WEF.org/agenda/2020/04/chatbots-covid-19-governance-improved-here-s-how/

The WEF sees the COVID-19 pandemic as an accelerator for chatbot technology, helping people around the world get more and more comfortable with accessing information this way. As we move beyond the pandemic, the adoption of chatbots in broader applications will continue to grow.

3.2 Potential for growth in AI/NLP services

According to an Accenture survey[20] 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[21] 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[22], 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.

3.3 Competitor and Market Analysis

Each of the commercial partners has provided an overview of competitor and market analysis for their sector. These are documented below.

3.3.1 Competitor Overview: Pilot I - Multilingual Knowledge Graphs for Knowledge Management across Sectors - 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

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

[21] Natural Language Processing Adoption Growth Insights, 2019

[22] 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

reach $11.74 billion by 2025[23]. 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[24].

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"[25] reports a recent shift in the metadata management market, away from 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

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

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

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

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"[26] competitors, SWC is ranked third for their approach to "active metadata management".

3.3.2 Market Trends: Pilot I - Semantic Web Company (SWC)

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.

To fulfil 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

[26] 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

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, analysing 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.

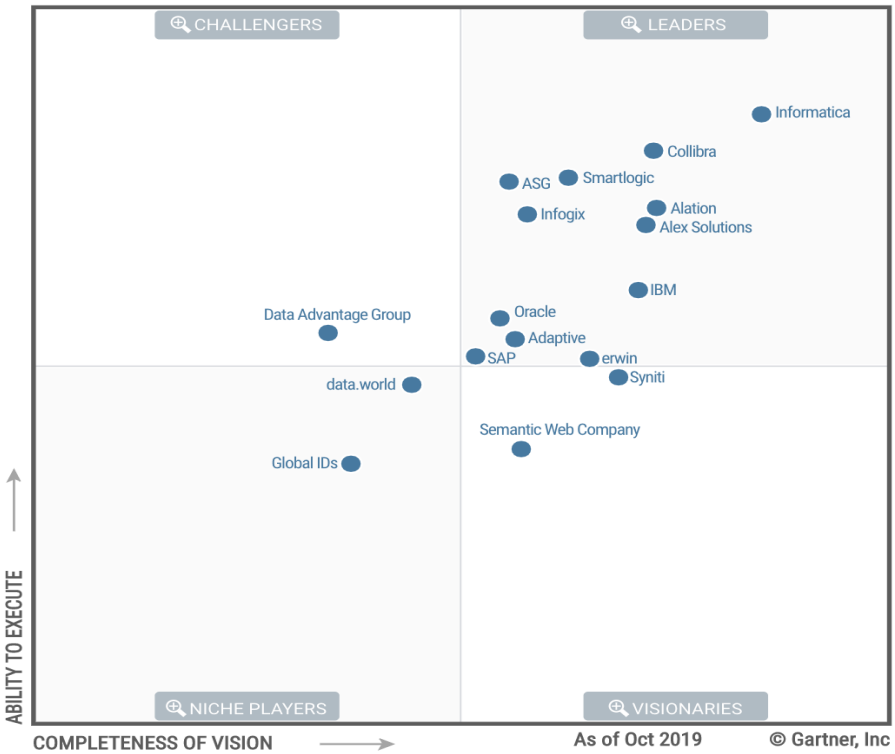


Figure 4: Magic Quadrant for Metadata Management Solutions

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 Fulfil 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.

3.3.3 Competitor Overview: Pilot II - Linking Lexical Knowledge to Facilitate Rapid Integration and Wider Application of Lexicographic Resources for Technology Companies - 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. New techniques like automated content creation through corpus analysis or machine translation, followed by postediting, show promise but have yet to achieve the level of quality produced by traditional methods; additionally, qualified lexicographers and editors cannot be recruited in all language combinations needed.

Market update

Text-based datasets are in increasing demand to support demand in the IT sector to support AI automation processes like speech recognition and text classification. The global market size was estimated at USD 956.5 million in 2019 and is expected to increase by 22.5% annually (CAGR) through 2027.[27] Amid this demand, language services companies are pivoting into services like data preparation and labelling, as well as into sale of parallel datasets derived from translation in specialized domains.[28] General structured lexical datasets that offer detailed labelling and

[27] https://www.grandviewresearch.com/industry-analysis/ai-training-dataset-market
[28] https://www.taus.net/insights/reports/language-data-for-ai-ld4ai

annotation also offer value for many use cases, especially for digitally under-resourced languages, and offering comparable, similarly structured datasets in multiple languages with the same data format facilitates integration and use for multilingual applications.

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.

In TIAD 2019 (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. The results of the 2020 TIAD were published jointly with the globalex proceedings.[29]

	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.

[29] https://www.aclweb.org/anthology/2020.globalex-1.0.pdf

A further line of work has gone 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*. It required a system to identify whether 2 definitions from the same lexeme in different dictionaries for the same language correspond to the same sense.[30]

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.

Having completed our pilots in 2020, OUP is looking beyond bilingual dictionaries to new linked lexical dataset opportunities. These include potential for automatically linking thesauruses to dictionaries at the level of meaning, and developing a suite of multilingual parallel sense annotated sentences linked to monolingual word-senses in dictionaries.

3.3.4 Market Analysis: Pilot II - Oxford University Press (OUP)

Market segments

OUP investigated the following market segments:

- Language services;
    - Machine translation (prioritised)

[30] https://www.aclweb.org/anthology/2020.globalex-1.0.pdf

    - 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

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

Derilinx' pilot project focusses on improvements to public services provided by government. The Derilinx pilot will provide:

i. access to public service information via a chatbot
ii. enhanced capabilities to find and compare open data from an open data platform using a dashboard

Both of these sub pilot projects provide support for multiple languages. The plan is for the functionality developed through these pilots to be integrated into the Derilinx datAdore product and so into open data platforms.

DatAdore is Derilinx' CKAN[31]-based data-sharing platform and their main product. In recent reviews with existing Derilinx customers, enhanced search was identified as a key feature that they would like to see improved. Table 2, below, shows an excerpt from Derilinx' analysis of closest competitor's open data systems for Public Service and Governments, focussing on the search functionality. This shows that adding a natural language search function to the datAdore product would give Derilinx a significant market advantage.

[31] https://ckan.org/

Company/ Product	Focus (Public Service (P) or Enterprise E)	Deployments	Pricing	Deployment Options	Search			
Feature					Metadata Search	Data Search	Geospatial Search	Natural Language Search
Description						for patterns in data		
DatAdore (Derilinx)	P & E	international	$$	hosted	yes	Roadmapped	currently CKAN standard, improved functionality roadmapped	
Datopian	Mostly P- some E	international	$$$/month	hosted/cloud	yes		CKAN Standard	
OpenDataSoft	P & E	international	pay as you go subscription hosting + volume + usage	hosted	facetted search			
LinkDigital	Mainly P /expanding into E	Australia centric, expanding		hosted	CKAN Standard		CKAN Standard	
Socrata	P	US Centric	$$$$/year	cloud	yes	maybe ??	yes	

Table 2: Derilinx competitor analysis for Search functionality

Derilinx plan to collaborate with Bielefeld University to improve the accuracy of their chatbot. This collaboration has been delayed due to availability of both parties but is now planned for the second half of 2021. Bielefeld have developed a system for generating a comprehensive set of questions from a knowledge-base, and the plan is to use this system to generate all potential questions for the chatbot.

As the user starts to type in their query, a list of suggested questions will be generated from the system, from which the user can select the most appropriate question. This will very much improve the accuracy of the results from the bot, as the user will only be presented with questions within the scope of the knowledge base.

This framework is ideal for the purpose of the Derilinx chatbot, which is designed to cover a particular knowledge-base, in the realm of public services. The bot will continue to provide services in multiple languages, in the first instance, English and Spanish, although this can easily be expanded to more languages.

This approach can also be used to enhance the search functionality within the datAdore product, where there is a known set of resources, and, in conjunction with the enhancements to general search, will provide Derilinx with a competitive advantage over its nearest rivals.

3.3.6 Market Analysis: Pilot III - Derilinx

The Derilinx open data dashboard pilot has been modified to access the European Open Data portal32. The European Open Data portal collects data from all national open data portals of European countries, to allow simplified access to data from all European countries. On the portal, some translations of dataset titles are available, but this is not consistent. The Derilinx pilot provides consistent translation of all dataset titles and metadata, along with translations for the data headings inside the datasets. As the dashboard allows the user to select two datasets, this enables the comparison of datasets from different countries, that were originally presented in different languages.

Chatbots during the pandemic

During the coronavirus crisis, chatbots have been used extensively for 24/7 health information dissemination. Advances in Natural Language Processing have meant that many consumers are familiar with the likes of Siri, Alexa and Google Home, and so have felt comfortable using chatbots to access critical information during the pandemic.

[32] https://data.europa.eu/data/

Adoption of chatbots for the provision of COVID-19 information can reduce the burden on healthcare call centres. In particular, chatbots can offer information customized to the needs and symptoms of the individual. Response to specific questions can be provided in an interactive manner, from a curated local source, which can include local guidelines and regulations.

According to the World Economic Forum33 (WEF), the COVID-19 pandemic has helped widen global usage for chatbot technology. The two most authoritative voices of the pandemic, the World Health Organisation (WHO) and Centre for Disease Control, have included chatbots in their websites to provide up-to-date, accurate information to billions on the spread of the disease and its symptoms. In April 2020, the WHO released a Facebook Messenger version of its Health Alert platform, specifically to counter COVID-19 misinformation34. Many governments are also launching chatbots to provide validated information to their citizens. Some healthcare companies modified their existing, trusted virtual assistants to adapt to the particular circumstances of the pandemic35.

The COVID-19 pandemic has acted as an accelerator for chatbot technology, allowing people around the world to become more comfortable in using this tool for all applications, driven by their use in healthcare.

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 English36. 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.

---

[33] https://www.WEF.org/agenda/2020/04/chatbots-covid-19-governance-improved-here-s-how/

[34] https://www.who.int/news-room/feature-stories/detail/who-launches-a-chatbot-powered-facebook-messenger-to-combat-covid-19-misinformation

[35] https://healthitanalytics.com/news/using-an-ai-powered-chatbot-to-meet-patient-needs-during-covid-19

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

Similar analysis of the figures for England and Wales[37] 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
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

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

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.

3.3.7 Competitor Overview: Pilot IV - Multilingual Text Analytics for Generating Real-World Evidence in the Pharmaceutical Domain - 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[38], 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

[38] Health Economics and Outcomes Research

B) Platform Providers

Company	Focus	Business Model	Life Science Exclusivity	Use Cases	Data Sources	Multi-linguality
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	Multi-linguality
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 is that insights can be delivered at quality levels that life science companies are used to from consulting agencies such as Cello Health and Kantar, while extracting them from non-traditional and unstructured data sources at machine speed and at unrivalled quality.

3.3.8 Market Analysis: Pilot IV - Semalytix

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[39], the interest in NLP is on the rise across many verticals[40], 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[41], 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.

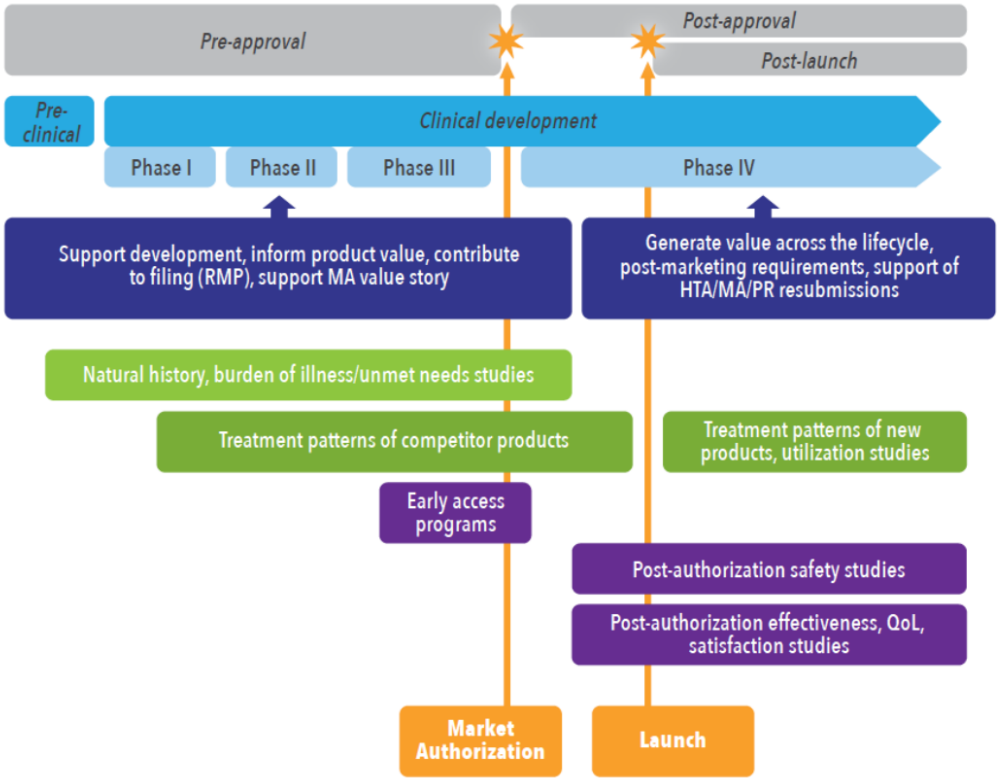
The quoted high-impact data sources are precisely the ones which hold the strongest potential for generating real-world evidence, i.e., assessing the value of drugs and medical interventions outside the controlled conditions of clinical trials. In particular, the interest in clinical narratives, either 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[42].

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

[40] This is supported by latest figures from Slator which quantify the addressable Language Technology market as reaching a volume of USD 23.8bn globally in 2020. The market volume remained stable despite significant disruption in the first half of the year due to the COVID-19 pandemic (https://slator.com/data-research/slator-2021-language-industry-market-report/).

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

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



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[43] 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®.

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

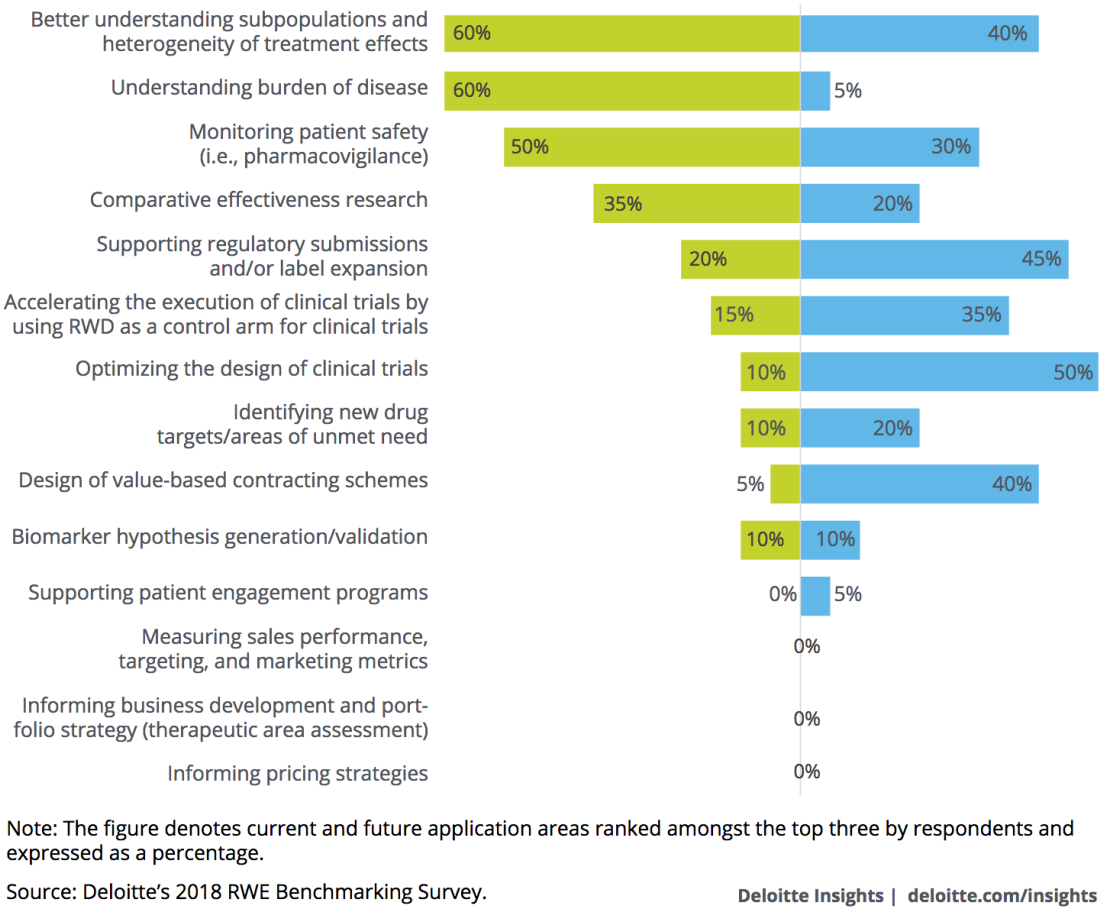


Figure 7: Perceived current (green) to future (blue) impact of RWE

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".[44] 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.[45] 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.[46] Based on these cases and further evidence, Olson (2019) concludes that the economic potential of RWE amounts to

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

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

[46] 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

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.[47]

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[48]).

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

[47] 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

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

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

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.

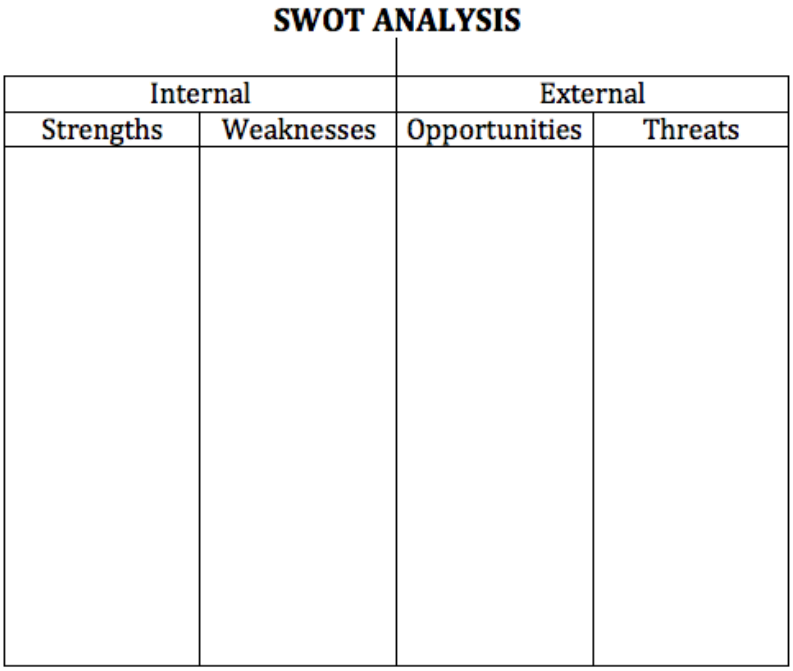


Figure 8: Swot Analysis Model

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.

4.1 SWOT Analysis for Prêt-à-LLOD

The analysis of the context of the Prêt-à-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 • Commercial 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 commercial partners' product offerings • Several pilots involve domain or task adaptation workflows which often make a crucial difference in practical usability for business	• Lack of marketing support relative to commercial competitors • Measurable uptake of the project components will be after project completion • Project extension increases risk around people turnover in commercial sector

use cases, thus increasing the likelihood of customer uptake (compared to offering rather "generic" LRs[50] or services) • Containerised deliverables preferred by market[51]	
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] • Greater investment post-COVID in language-based AI • Greater acceptance of AI through COVID applications (e.g. chatbots)	• Similar product/services offered by competitors • Potential entry barriers: companies may have their own workflows in place already, may enable some market problems to be resolved through alternative means

[50] Language Resources

[51] https://www.european-language-grid.eu/wp-content/uploads/2021/02/ELG-Deliverable-D7.8-final.pdf

[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.

5 Collaboration with academic partners

The commercial partners have collaborated with the academic partners in the project, resulting in improvements to the pilot projects, as well as, in some cases, giving the academic partners an opportunity to use the pilots as a testbed for their developments.

	NUIG	UZAR	UNIBI	GUF	DFKI
OUP	1	1,2			
SEM		3		3	
SWC				4	4
DLX			5	6	

The academic partners are:

DFKI Deutsches Forschungszentrum für Künstliche Intelligenz

GUF Goethe Universität Frankfurt

NUIG National University of Ireland, Galway

UNIBI Universität Bielefeld

UZAR Universidad de Zaragoza

Details of the Collaborations

1 OUP has explored the possibility of combining different approaches to cross-dictionary linking to generate bilingual outputs.

2 OUP and UZAR have shared their approaches to modelling uncertainty in dataset links and both partners have refined their models based on this information.

3 SEM will be exploiting resources and workflows provided by GUF and UNIZAR in order to inform model transfer across languages.

4 SWC is attempting to create a solution for generating inflections of compound terms using resources developed by DFKI and GUF.

5 DLX will be applying the methodology developed by UNIBI to generate questions and answers for any set of text to their chatbot.

6 DLX is considering the incorporation of the GUF-developed tool to convert any file to RDF into their pilot.

6 Conclusion

This second Business Development Report has given the commercial partners an opportunity to elaborate on and evaluate the business strategy for their pilots. Since the previous version of this report, each partner has learnt from their intermediate results and adjusted their plans as necessary.

Market changes, as a result of COVID in particular, have had a significant impact and the SWOT and Competitor Analyses have been updated accordingly. As the technologies and standards to be delivered by other work packages in the Prêt-à-LLOD project become available, they are being incorporated into the pilots. The benchmarks and market research highlight the importance of collaboration between the commercial and academic partners in the project.

The next 12 months will put the focus on the exploitation of the pilot projects, with the partners receiving feedback from real customers and market validation of the anticipated commercial opportunities.

7 Appendix – References

TIAD 2019 (Gracia at al. 2019), https://tiad2019.unizar.es/

Kumiko Tanaka and Kyoji Umemura. 1994. Construction of a bilingual dictionary intermediated by a third language. In Proceedings of the 15th Conference on Computational Linguistics, Volume 1. Association for Computational Linguistics, 297--303.