Building the Legal Knowledge Graph for Smart Compliance Services in Multilingual Europe

D5.7 Demonstrator for pilot 3

PROJECT ACRONYM	Lynx
PROJECT TITLE	Building the Legal Knowledge Graph for Smart Compliance Services in Multilingual Europe
GRANT AGREEMENT	H2020-780602
FUNDING SCHEME	ICT-14-2017 - Innovation Action (IA)
STARTING DATE (DURATION)	01/12/2017 (40 months)
PROJECT WEBSITE	http://lynx-project.eu
COORDINATOR	Elena Montiel-Ponsoda (UPM)
RESPONSIBLE AUTHORS	Pascual Boil (CUATRECASAS), Elsa Gómez (CUATRECASAS), Pablo Calleja (UPM)
CONTRIBUTORS	Elena Montiel-Ponsoda (UPM), Patricia Martín-Chozas (UPM)
REVIEWERS	Pieter Verhoeven (DNV.GL), Christian Sageder (Cybly)
VERSION | STATUS	V1.0 / Final
NATURE	Demonstrator
DISSEMINATION LEVEL	Public
DOCUMENT DOI	10.5281/zenodo.4300691
DATE	30/11/2020 (M36)

This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 780602

VERSION	MODIFICATION(S)	DATE	AUTHOR(S)
01	Initial draft	11/11/20	Pascual Boil (CUATRECASAS)
02	First completed reviewed version including comments and recommendations from reviewers	23/11/20	Pascual Boil, Elsa Gómez (CUATRECASAS)
03	Comments and suggestions provided by reviewers	25/11/20	Pieter Verhoeven (DNV.GL), Christian Sageder (Cybly)
04	Full review and content redistribution proposal	26/11/20	Pablo Calleja (UPM), Elena Montiel-Ponsoda (UPM), Víctor Rodríguez-Doncel (UPM)
05	Additional content and review of new sections	27/11/20	Pablo Calleja (UPM), Patricia Martín-Chozas (UPM)
1.0	Final version adjustments	30/11/20	Pascual Boil (CUATRECASAS)

ACRONYMS LIST

AD:	Active Directory
API:	Application Program Interface
BOE:	Boletín Oficial del Estado
DBMS:	Data Base Management System
EntEx:	Entity Extraction
IT:	Information Technologies
KM:	Knowledge Management
LATAM:	Latin America
LKG:	Legal Knowledge Graph
M&A:	Mergers and Acquisitions
MT:	Machine Translation
MVP:	Minimum Viable Product
NLP:	Natural Language Processing
NMT:	Neural Machine Translation
PPT:	PowerPoint presentation
QA / Q&A:	Question and Answer
QADoc:	Question Answering from Document
SME:	Small Medium Enterprise
SSO:	Single-Sign-On
StrEx:	Document Structure Extraction
TimEx:	Time Expressions

EXECUTIVE SUMMARY

This deliverable summarizes the final results of the Lynx use case developed for labour law and explains all relevant parts of the pilot and activities done around it (business use case into the Lynx project).

Section 1 introduces some of the difficulties that lawyers or legal experts encounter when working with legal documents from different jurisdictions in multiple languages. In this use case, we aim to demonstrate how semantic technologies and the use of knowledge graphs can contribute to assist lawyers (and non-lawyers) in accessing and processing international legislation. How to apply these technologies and how to simplify the interaction of users with legal data by providing an easy-to-use interface, are some of the challenging aspects that we face in this pilot.

Section 2 describes our use case in more detail, focusing on the main objectives of a law firm like Cuatrecasas, and how the solution built in Lynx can help its clients and prospects. We explain our envisioned solution by providing detailed information of the functional requirements derived from several functional meetings. At this point, a level-of-implementation gap analysis is also included to detail and explain what you can find finally implemented in this last deliverable.

As part of this final demonstrator, in Section 3 we show the core of the solution, also known as the Minimum Viable Product (MVP). Here we explain scope redefinitions and design changes that differ from the initial idea. In this section we want to highlight those parts of the project which have been fundamental within the pilot, even if not all of them have finally materialized in the final solution. To better illustrate the resultant application, and how it works, we have included a Tour Guide through the Beta version of this implemented MVP with commented screenshots, introducing the main functionalities.

In Section 4, we explain the different components of the Application Architecture (functional but also giving some technical details). This includes a description of the core architecture and main components and how they interact with the Lynx platform and the Lynx services as “building blocks”.

To conclude, Section 5 explains relevant activities and experiments, intermediate lines of work, that have also been important for the project, even if they have not materialized as part of the final version of the software. Additionally, an outlook and next steps are described in Section 6.

The report is supplemented with four annexes, which provide additional information and context:

I. “Sample of real questionnaires”, based on previous real international operations by the company
II. “Legal Knowledge Graph in the labour use case”, putting our use case in the context of the global LKG
III. “Chatbots, QA and Semantic Search working together”, an analysis of these different technologies and how and why to combine them.
IV. “User validation and feedback”, as an example of the work done on results and accuracy testing and training services (QADoc in this case)

TABLE OF CONTENTS

TABLE OF FIGURES

1 INTRODUCTION

Companies need to comply to different regulations. Almost all of them are published in local sources (usually public institutions at any territory level), and most of them are only available in their official local languages. This problem is acuter at European level: although there is a common regulation and regulatory framework, the extent to which European directives have been transposed can differ greatly.

Cuatrecasas is a full-service law firm, which, although leading in Spain and the Iberian market, also provides global legal advice to international companies. As many of our clients are international companies, we deal with many languages and country laws and regulations.

Our pilot focuses on labour law, which typically involves several international operations because of our clients' geographical expansion (e.g., mergers & acquisitions and due diligence). In large corporations, geographical expansion and differing workers' rights are a common problem, as the regulations of each country differ. Other large companies, although not international, can still face the same problems with sectorial or geographical national agreements.

This labour law use case can be extended to other legal practices like tax, intellectual property rights (IP) or data privacy and personal data (recently regulated in the GDPR directive at European level but with global impact). The problem regarding cross-border regulations is increasingly common in a globalized economy, where the level of regulation, the number of laws and the frequent changes they undergo are also increasing yearly.

In this project, we aim to cover two use cases that have no significant functional differences between them. The first one is targeted at Cuatrecasas lawyers ("the internal use case") to enable a more efficient access to legislation across jurisdictions; and the second use case ("the external use case"), intended at Cuatrecasas clients, providing their internal legal department teams or even their human resources department, a direct and secure access to legislation.

Summary of the two use cases:

(1) **Support tool for our internal lawyers**, to advise clients on their international businesses, as well as international mergers and acquisitions (M&A) that commonly involve several law firms specialized in their local laws.

(2) **External service for our clients to be directly use without intermediaries** (global organizations)

We envision the same technical solution for both use cases with minor differences between them. Our solution resembles a Legal Chatbot in the sense that it relies on a user interface specially thought for non-legal experts (non-lawyers, simply junior lawyers or paralegals) in which they can formulate a full question in natural language.

The final global solution can be rather defined as a "smart" natural language search tool for lawyers, where results are texts or excerpts directly extracted from the law ("technical" legal language). However, a chatbot-like interface offers the ideal interaction scenario for non-legal experts, relying on a Question Answering (Q&A) system that simplifies the access to regulatory sources and helps them interpret the legal content. Combining the semantic search and Q&A chatbot-like interface in the same application will be one of our main challenges. (Additional information about this topic is available in Annex 4).

In recent years, the legal sector has sought to use user-centric technology to democratize access to justice. For us, legal chatbots have the potential to open access to justice for everyone. For this reason, in Cuatrecasas, we have decided to research and develop a pilot test inspired by legal chatbots as "legal

assistants" to evaluate the latest technology; and how it can help improve our legal services and the provision of legal advice to our clients more efficiently.

Relying on our previous experience with chatbot platforms (Microsoft Azure bot services and Google cloud with DialogFlow), we know that they tend to focus on Q&A models, which require to put huge human effort on manually identifying precise questions and answers. This model is valid and probably efficient for many limited-scope knowledge environments whose content does not change frequently. However, it is not the ideal solution when addressing "laws" in multiple countries and languages, due to the quantity and variety of regulations in the world and the continuous changes they undergo. Trying to cover all potential legislation queries through a predefined list of question and answers is not realistic and, for sure, not efficient in economics terms.

Because of the abovementioned limitations, in the Lynx project we devise a solution based on semantic search technologies combined with a Q&A system trained with Machine Learning techniques on a specific domain, the legal domain, and more specifically, on the labour law sub-area. The semantic search part is key for the future success and sustainability of this use case. In this regard, we assume that more than 80% of the simple legal questions are directly covered by semantic search, and without investing a huge amount of time on training effort by experts. Our underlying aim is to only dedicate additional resources to work on specific tailored-made question-answers if our firm can deliver additional value by "interpreting" the law for our clients.

2 USE CASE DESCRIPTION AND OBJECTIVES

2.1 OBJECTIVES OF THE PILOT

With the labour law pilot, we aim to achieve two main objectives:

- To provide users with potential answers related to his/her legal questions involving several jurisdictions and natural languages. We regard users as legal experts that are to interpret the results provided by the system in his/her own language (we do not expect the system to replace the legal expert).
- To improve efficiency in searching and accessing legal documents from different jurisdictions and contribute to an enhanced understanding of the results (and facilitate comparison between jurisdictions). To enhance our company solutions to better position the company in the legal market, not only as an innovative, but also as a more global law firm.

Achieving these purposes will translate into an increase of efficiency in task performance in both internal processes of the law firm and external interactions with our clients.

For the second year in a row, Cuatrecasas has received the Financial Times award for the most innovative law firm in Europe. This project is yet another opportunity to enhance our existing legal services through the use of technology, and even to generate new businesses through the delivery of legal products that really add value to our customers. With this pilot our objective is to demonstrate that this is a feasible and efficient solution. The pilot will cover a subset of European languages, (namely, English, Spanish, German and Dutch) and jurisdictions (Europe, Spain, Germany, Austria and The Netherlands). Additional mechanisms should be put in place so that further languages and jurisdictions are considered according to Cuatrecasas' client's needs (frequently involving non-EU countries, which are out of the scope of this project).

As a pilot, we should be able to evaluate several technological and business aspects:

Technology:

- Evaluating the current status of neural machine translation trained with highly specialized texts from the legal domain
- Testing semantic technologies (NLP, Semantic Search, …) and contributing towards improving the current Q&A/chatbot solutions to build a custom solution that can be almost self-maintained.
- Checking the viability and sustainability of the Legal Knowledge Graph as part of the Lynx platform; finding an independent way of having open access to legal resources, ensuring that it is a service that can be provided in a real business environment.

Business:

- Assessing the level of accuracy of the retrieved answers and the time it saves to a lawyer who is not an expert in the jurisdiction involved. As an internal tool to be used in international M&A operations dealing with legal multi-country questionnaires, this would be very useful for those jurisdictions in which we do not have specialized lawyers.
- Testing the pilot with current and potential customers and identifying their willingness to pay for this additional Cuatrecasas service. A market analysis should be foreseen to find out which other segments (sectors and parties) could be also interested in this solution (e.g., SME's, individual lawyers and small law firms).

2.2 INTERNAL USE (CUATRECASAS LAWYERS)

2.2.1 General description and functional overview

As a support tool for Cuatrecasas' lawyers this application should be executed internally (inside the corporate Cuatrecasas network) where the users are identified (automatically) into the Cuatrecasas domain. At this stage, implementing an additional user authentication interface is not necessary.

Cuatrecasas provides a range of services to clients, including (i) specific operations, which typically involve a project with a limited scope and period; and (ii) general legal advice, which is usually categorized by practice area (e.g., labour, tax and corporate) due to the different legal specializations required. Also, our lawyers may work on more than one matter, as well as for multiple clients, at a given time.

For this reason, the system must provide lawyers with the tools they need to organize and optimize their tasks, enabling them to configure and save their favorite options (more common/default): personal or client/company.

Although Cuatrecasas has offices in multiple countries, the firm's official languages are Spanish, English and Portuguese. Despite our specialization in the Spanish and Portuguese jurisdictions, we now offer global international coverage to our clients, with a focus on LATAM (Latin America). Our typical clients are big (Spanish and Portuguese) companies with business around the world, as well as international companies with subsidiaries or business interest in Iberia or LATAM.

Countries usually publish their laws in their own official languages. The main problem non-local lawyers usually face is accessing and understanding foreign local laws and regulations, which are not often available in other languages.

For this internal use case, we assume that our users are legal experts. Often, they are junior lawyers who are tasked to investigate external regulations. Currently, these lawyers have to contact our internal Knowledge and Innovation Team to find out about (i) the legal particularities of a specific country/jurisdiction; (ii) the legal sources available; and (iii) whether we count on local lawyers from partnering institutions we can contact, if necessary. These lawyers are accustomed to use legal databases and other information resources (e.g., the ones provided by LexisNexis, Thomson Reuters and vLEX). Moreover, they usually have a good command of the legal terminology in their own language and in English, but only very limited knowledge of the legal terminology in other languages.

The aim of this use case is be able answer several legal questions about the labour law, in several countries-jurisdictions. Cuatrecasas as a law firm, with international coverage frequently participates in international cross-border operations (typically international M&A, sometimes leaded by us as a prime/leading law firm, sometimes only being part of the operation as a local law firm because our specialization in Spain and Portugal), and as part of these operations there are specific legal subject questionaries (one of the most frequent and complex is the part related to the labour aspects on each country) where several key aspects of the operation (company acquisition or simply geographical expansion as new factories or new offices) must be analyzed, based on the specific country regulations. In this typical use case, one of the most representative for this Lynx pilot, we receive labour law questionaries typically in Excel file format (an extract from a real example questionary is shown in Annex 1) and our labour lawyers must answer them, filling in the document with the different countries-jurisdictions considerations, and referencing to the local legislation.

We are envisioning a system/application where the user formulates a complete query regarding labour law and workers' regulations, specifying one or more jurisdictions, and the system returns the most relevant information based on the direct texts of the law, including the following:

- The most precise answer possible (when the question is specific, asking for a value, data and name).

- The paragraph(s) related with the topic/question, where the possible answer appears as part of the text [ideally highlighted].

- The context by showing the article (and section) from which the paragraph(s) is extracted, showing the number and title, and allowing the user to view the full text of the article and law, which the user should be able to access and download.

Complex legal questions are almost impossible to answer by only highlighting parts of the law. Context and additional information are often needed. This additional information is sometimes difficult to incorporate into a question and these context words are not always easy to find directly mentioned in laws. **Our system is designed to be used as an intelligent search tool, providing legal guidance to lawyers**, to help substitute or minimize some of their less-value work.

2.2.2 Functional requirements

As part of the functional analysis and scope definition of the future application, in the table below we inventory the list of functional requirements. This list of requirements is prioritized and grouped by functional modules:

- Login Module [LOGIN]
- Configuration Module [CONFIG/DEFAULTS]
- Question Answering Module [Q&A]

We prioritized these modules based on meetings with final users, where priority was given to information that is key to specifying the Minimum Viable Product (MVP) with the basic and valuable functionality. The implementation of the modules is described in sections 3 and 4. Adjustments and changes to the initial modules are explained in section 5.2 “Troubleshooting and Lessons Learnt”.

Id & Module	Functional description	Priority/ Importance
REQ01 [Q&A]	The user should be able to write a legal (labour-related) question in all supported languages. The system will then return the possible answers based on the related law (selection of laws or regulatory documents from possible jurisdictions).	HIGH [MVP1]
REQ02 [Q&A]	The user will be able to select (1 to N) jurisdictions/countries to ask/work with (from the jurisdictions considered in this project).	HIGH [MVP1]
REQ03 [Q&A]	The system will be able to identify the language used in the question. If no other rule or personal setting is defined, the system should show the results in the user’s language.	MEDIUM-HIGH [MVP2]
REQ04 [CONFIG. /DEFAULTS]	The system will allow users to save personal settings, which should be default when they use the application. However, users should be able to change these options for specific questions. Some of the PERSONAL SETTINGS should include DEFAULT LANGUAGE (for answers), DEFAULT JURISDICTIONS and DEFAULT DOCUMENTS/LAWS.	MEDIUM-HIGH [MVP2]
REQ05 [CONFIG. /DEFAULTS]	Users can work with different clients (CLIENT SETTINGS). The system will allow users to save specific configuration/defaults for each client, as well as to change this configuration at any time when using the application. The priority rule to apply these default settings would be 1st CLIENT and 2nd PERSONAL. Therefore, if the client configuration is selected, it should be	LOW

	applied, and the personal settings will be selected automatically when applying the client settings is not possible.	
REQ06 [CONFIG. /DEFAULTS]	The Q&A execution is restricted to a regulatory set of documents (country laws and other regulatory/compliance documents). Users should be able to select (add or remove) their "subset of the Legal Knowledge Graph (LKG)" to perform their search/question-answer. This subset will be created by filtering the full LKG by Type of Document ("Law" by default), Country/Jurisdiction, Company (*), Sub-type of Law, Legal Domain/Specialization ("Labour" by default). (*) The default configuration will be the labour law of the selected jurisdictions [MVP].	MEDIUM-HIGH [MVP2]
REQ07 [CONFIG. /DEFAULTS]	Users should be able to upload their own documents (e.g., internal company agreements and sector-specific agreements) to the LKG (we assume that LKG will allow public and private documents, which will be secured and classified so they can be filtered). Document formats will be mainly PDF and MS Word (e.g., doc and docx).	LOW
REQ08 [Q&A]	The system should be able to show several types of answers: (1) Related paragraph(s) (direct result list) and the article and section they belong to, as well as the document (names/titles) [MVP]. access to each part of the document should also be allowed (i.e., article and section) [MEDIUM MVP]. (2) In these paragraphs, the system should highlight the text that most relates to the question [LOW]. (3) The system should show specific personalized answers (it will not be texts/parts of the law), and it will be based on pure question-answer rules. Specific and personalized answers will not always exist (hard manual work). However, when available, they must be the first results shown to the user (i.e., 'sponsored results') and they must be highlighted [LOW]. (4) The most precise answer should be shown first (when the question asks for a concrete value in the document (a date, a number, legal authority name, …) [MEDIUM]	MEDIUM-HIGH [MVP2]
REQ09 [Q&A]	The system will also return relevant (and recent) case law related (jurisprudence) to the question (by article–legal topic equivalent) in the different jurisdictions.	LOW
REQ10 [Q&A]	TEXT AND DOCUMENT (TRANSLATION) requirements. The user will be able to view: • Full document (ideally in its original format), and be able to print it out • Full Translated document • Complete article text (local language), as well as translated into the other languages • Paragraph text (local language), as well as text translated into the other languages • In all content, the minimum options for languages/translations should be (i) user own language, (ii) English, and iii) local document/jurisdiction language.	MEDIUM
REQ11 [Q&A]	The user will be able to mass upload questions from an Excel file (QUESTION LIST). The system should be able to show the results in the application, while also providing the option to generate the results in an Excel file.	HIGH [MVP2]
REQ12 [Q&A]	QUESTION HISTORY. The system will save all the results and interactions, and users will be able to navigate back from previous question-answers.	MEDIUM-HIGH [MVP2]
REQ13 [Q&A]	The system will allow users to export all or a selection of the question-answers in an Excel sheet.	LOW
REQ14 [Q&A]	Users could rate the answers (expert users) to be able to provide feedback and improve the internal algorithm.	LOW

REQ15 [Q&A]	Users should be able to mark answers as “favorites,” which they could use for future analysis.	LOW
REQ16 [Q&A]	Users will be able to COPY (to clipboard) the paragraph result so they can paste it in other documents (e.g., Excel, Word or emails).	LOW
REQ17 [LOGIN]	SSO. Internal users (Cuatrecasas) do not need to enter any additional usernames or passwords into the system. The application will need to be integrated into the Cuatrecasas authentication system (Microsoft AD).	HIGH [MVP1]

2.3 EXTERNAL USE CASE (DIRECT ACCESS FOR CLIENTS’ LEGAL DEPARTMENTS)

In this section we briefly refer to the adjustments that such an application should undergo if Cuatrecasas aimed at offering it directly as a product to its clients (or prospective clients), to be used by their legal departments.

The external version would differ from the internal version in the following ways:

- ARCHITECTURE: The solution would be accessible on the internet (outside of the Cuatrecasas network).

- SECURITY: The system should be highly secure, guaranteeing high levels of availability, security and privacy. Before going live externally, an external hacking service to test the system would be mandatory.

- LOGIN: Users would need a special identification (username and password) with a “remember password” feature.

- USER MANAGEMENT AND DELEGATE USER MAINTENANCE: The system should provide a mechanism to manage internal users from the client company, with different roles (client admin).

- One client could have access to different company configurations (1:N) (e.g., to big companies/groups by their different local companies or subsidiaries).

    - Q&A MODULE IMPROVEMENTS / ENHANCEMENTS: “Recommended answer” to refer to specific Cuatrecasas’ answers for common questions.
    - Chatbot interaction, with a “refine or supplement information” feature.
    - “Precise answer” Improve accuracy of the results and emphasize the importance of providing a “precise answer” in as much as possible.

- In addition, we could simplify/reduce certain features that are of less interest to users (e.g., RATING ANSWERS and QUESTION LIST —mass upload).

3 PILOT DESCRIPTION

3.1 INTRODUCTION

Once the objectives of the pilot have been introduced and the functionalities of the application described, this section presents the final delivered product (Cuatrecasas Lynx Application), and explain the main modules the project.

3.2 PILOT OVERVIEW

Figure 1 shows an overview of the Pilot and the different modules that it comprises.

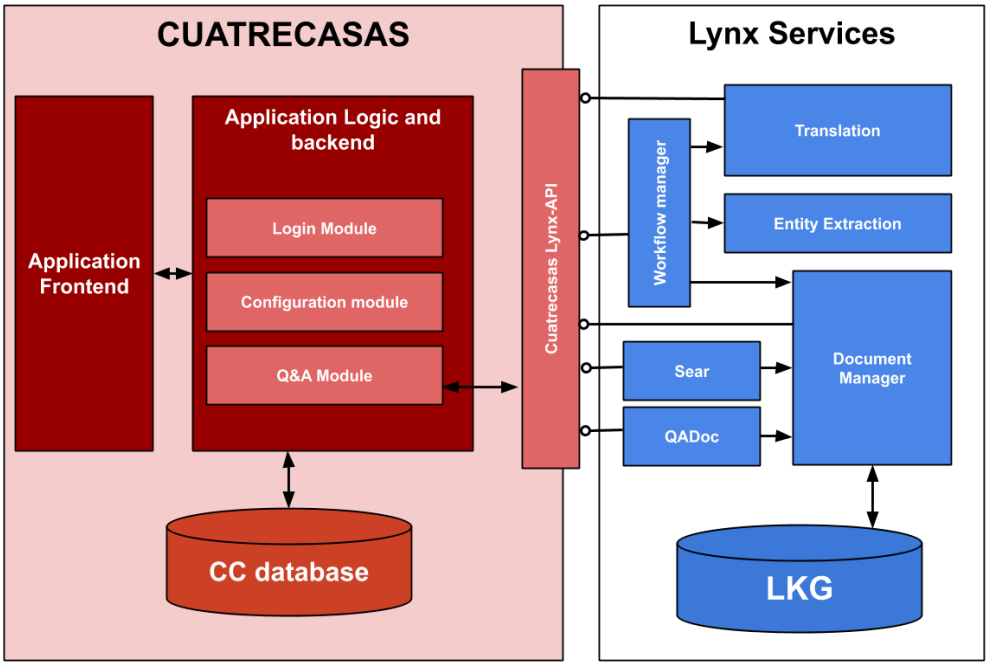


Figure 1. Pilot modules and components schema

Cuatrecasas Lynx Application is comprised of 4 main parts or components:

- **Front-end Application**, with the presentation layer responsible for the user experience
- **Back-end Application** and business logic layer, to encapsulate the different defined modules (login, configuration and Q&A modules) and provide all required application functionalities
- **Application database**, to ensure data persistency
- **Cuatrecasas-Lynx API**, a middleware component to encapsulate and centralize interaction with Lynx Services

The Lynx services used in the Cuatrecasas Lynx Application are described in the following:

- **SEAR Search Service (UPM and Cybly)**

  Two SEAR services are used in order to retrieve documents from collections previously defined by end users.

  Lynx Standard Search Service from Cybly is used to select documents from the LKG/DM based on metadata and content filters.

  An advanced UPM Search Service has been finally used as enhanced service to generate the answers as query results. The service generates a first list of ranked candidate answers (previously broken

down into paragraphs), the service also highlights the text segment that is responsible for the selection. The SEAR service uses the document enrichment processes performed by the Lynx annotation services to allow filtering out searches and to score the results based on the question. Additionally, this service uses Query Expansion (QE) mechanisms to improve search precision and cover the main use case requirements.

Both SEAR services are based on Elasticsearch technologies, but each of them provides different search parameters. Right now, several tests are being performed to assess the quality of the results. More information about the SEAR service will be available in D5.8 by month 40.

- **QADoc Service (SWC)**

The QADoc service receives a question in natural language and a source text to find a precise answer within it. Only when the service returns a result with a high level of confidence, the application will show this result to the user.

The use and content of this service has changed a lot throughout the pilot. Initially, as it was the initial architecture design, all the answers received by the end user should came directly from the QADoc service (first, the SEARch service preselected those parts of the document where the answer could most likely appear and the QADoc service returned only those where it found a potential "precise" answer to the question) but today the most important part of the use case resolution is done directly by the UPM Search service which is retrieving all paragraphs results and QADoc is only used as a next complementary step to present additional “precise answer” on it.

This service has been developed as part of Task 3.4 (see D3.4 for more details).

- **TimEx Service (UPM)**

Time expressions are very relevant in any legal document. For example, expressions for deadlines or regulated procedures are common in the labour context, such as “something has to be done 10 days after the contract is signed,” “the probationary period does not exceed six months,” or “the cost of dismissing an employee is 20 days per worked day.” Therefore, the **TimEx service** is of particularly relevance in this use case to identify the time expressions mentioned in documents. Nowadays, this service is used to provide annotations in the document ingestion process, but it will be used in a new future advanced version of the QADoc.

- **Machine Translation Service (TILDE)**

The translation service provides automated machine translation by using the Tilde MT cloud platform. Currently, the translation service provides support for a runtime scenario and an endpoint for the Lynx platform asynchronous process in the background. Neural Machine Translation (NMT) systems were trained for the language pairs selected by the demonstrator in the framework of WP3. In the domain of labour law, specific legal and business data was gathered and processed before training the NMT systems on a mix of broad-domain and in-domain data to be able to translate both in-domain and out-of-domain texts. For more information, see deliverable D3.2 Intermediate translation services.

- **Document Manager (DCM) and the LKG**

The document manager (also referred to as DCM) is a central part of the Lynx platform, as it is where the documents are stored and maintained. Its basic functions include storing documents and their annotations, particularly regarding maintaining their synchronization, providing read and write access, as well as updates of documents and annotations. The document manager can be queried in terms of annotations (e.g., “which documents mention this entity?”), as well as in terms of documents

(e.g., "what are the contents/annotations of document X?"). The interface includes a set of APIs to manage the following resources within the Lynx platform: collections, documents and annotations. DCM is responsible for storing the Legal Knowledge Graph (LKG) and the documents once they have been processed through the different workflows. For more information, see WP4 documentation, deliverable D4.4.

- **Workflow Manager**

The Workflow Manager is responsible for the effective orchestration of the micro-services to carry out workflows. Workflows are combinations of both parallel and sequential tasks, which are specified using Directed Acyclic Graphs. For more information, see WP4 documentation, deliverable D4.4.

This "macro service" also uses other Lynx Services, namely, Entity Extraction (EntEx) and Document Structure Extraction (StrEx).

3.3 MAIN CHANGES FROM MVP ALPHA VERSION TO CURRENT MVP (BETA VERSION)

In this section, we present the main improvements from the Alpha version presented in D5.3 to the current Beta version.

Some of the improvements and changes with respect to the alpha version are specified in the following:

- **From one single document repository limited to a single language (English) to several documents' repository with the different supported languages in Lynx (English, Spanish, Austrian, German, …).** As QADoc service has only been trained in English for the legal domain, the documents in the application are translated and its English translation is used for different services.
- **Internal access for Cuatrecasas users without SSO to allow internal SSO validation and external access out of the Cuatrecasas network and directory.** While the Lynx platform performance, accuracy and exploitation is being discussed and assessed, further tasks and features are being considered as next steps to test the use of this pilot with some of Cuatrecasas clients
- **Translation to any supported Lynx language at questions and answer level.** The limitation of using only Spanish and English languages is removed. Also, at this point we have achieved several levels of accuracy regarding specific domain-specific training effort (English, Spanish, German, Austrian, Dutch …)
- **Queries can obtain results from documents of several jurisdictions,** not only one (Spanish).

Also, additional functionalities have been included:

- Personal/user defaults
  - Capacity to upload, ingest new documents into the system using the Lynx platform: LKG, Workflow with the Annotation Services and the Document Manager
  - Allows users to create ad-hoc personal document repositories (corpus) to personalize their Question Answering environment.
  - Manage preferred languages (to present the results in that language) and pre-selected jurisdictions (to group results and provide comparisons between them)
- Securitization for companies:
  - Restrict access to Cuatrecasas private documents by other LKG users, pilots and future use cases
  - Restrict access to some companies' information by non-granted users (internal or external)

3.4 VISUALIZATION OF CUATRECASAS APPLICATION (BETA)

Authentication

When accessing the application, the internal user ("ELGD" in the example) is recognized and auto validated through a SSO mechanism that integrates with Cuatrecasas' internal AD, skipping the login screen (**REQ17**). For any non-Cuatrecasas user, an identification screen with user and password is needed as Figure 2 shows (**REQ01**).

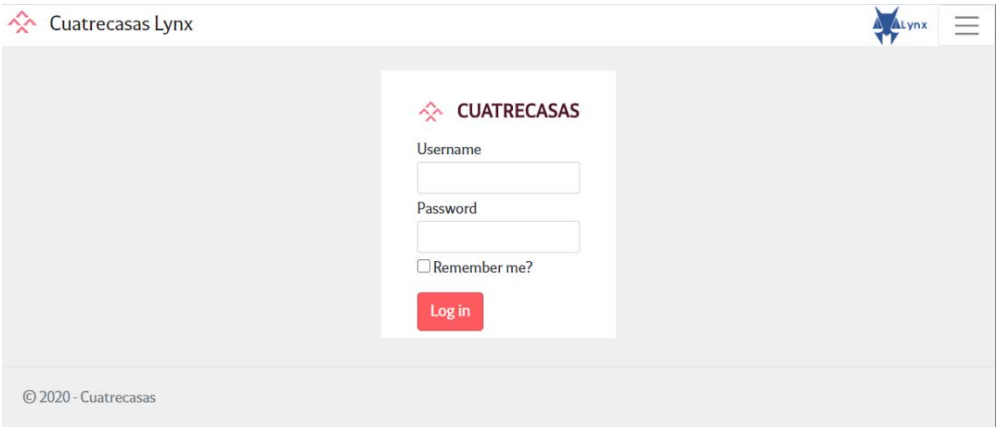


Figure 2. Screenshot External user login screen

Any new user has access by default to a document repository. The main document of the repository is "Workers Statute in Spanish, consolidated version". Users have the capability to create their own favorite document sets through the "User profile" and "Company settings" sections within the top menu.

User profile

Any individual user has a "User profile" configuration in which he/she can specify some personal defaults (**REQ04**):

- Language: The output language of the application results
- Jurisdiction: The preferred jurisdictions to work. Each result content found related to these selected jurisdictions will be splitted into different tabs, into que "Question" section

These two parameters are mandatory for any new user. Language as "Spanish" and Jurisdiction as "Spain" are the default values for any new user into the system.

Another important feature in this section is the **personal documents defaults**. Each user can specify individual collections of documents. These collections are specified as "favorite" and will be selected to be used as a corpus repository at question level (Figure 3).

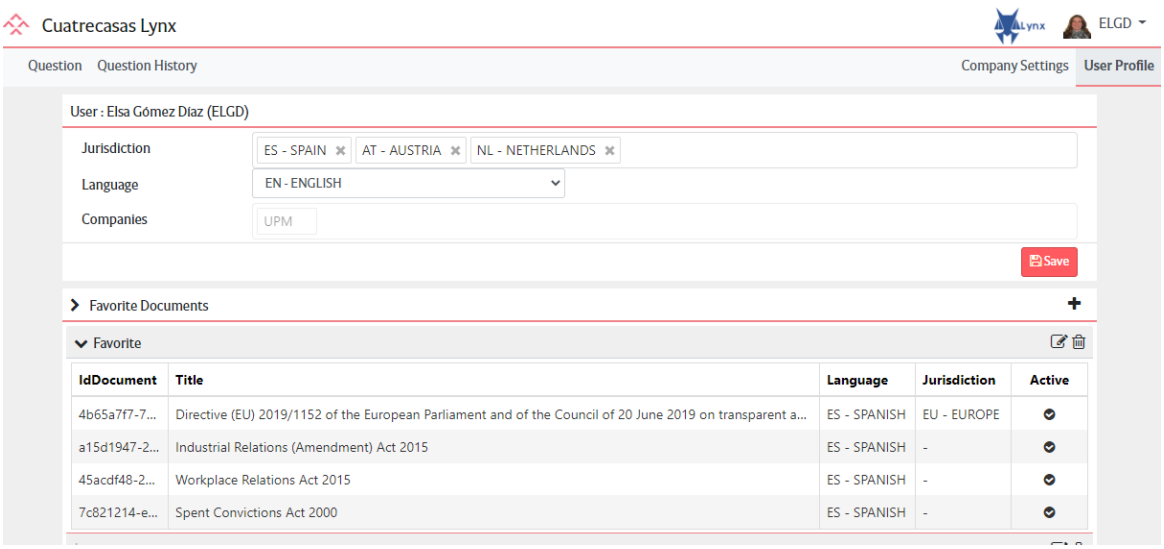


Figure 3. Screenshot User profile favourite Documents

The document selection is done through a filtering option that retrieves documents from the Lynx repository (DM and LKG) based on security rules, that restrict physical and logical access rights to any document stored. From the user point of view, we have segmented LKG content for CC use case into 4 main document types ("Legislation", "Collective Agreement", "Company Agreement" and "Case law") (see Figure 4).

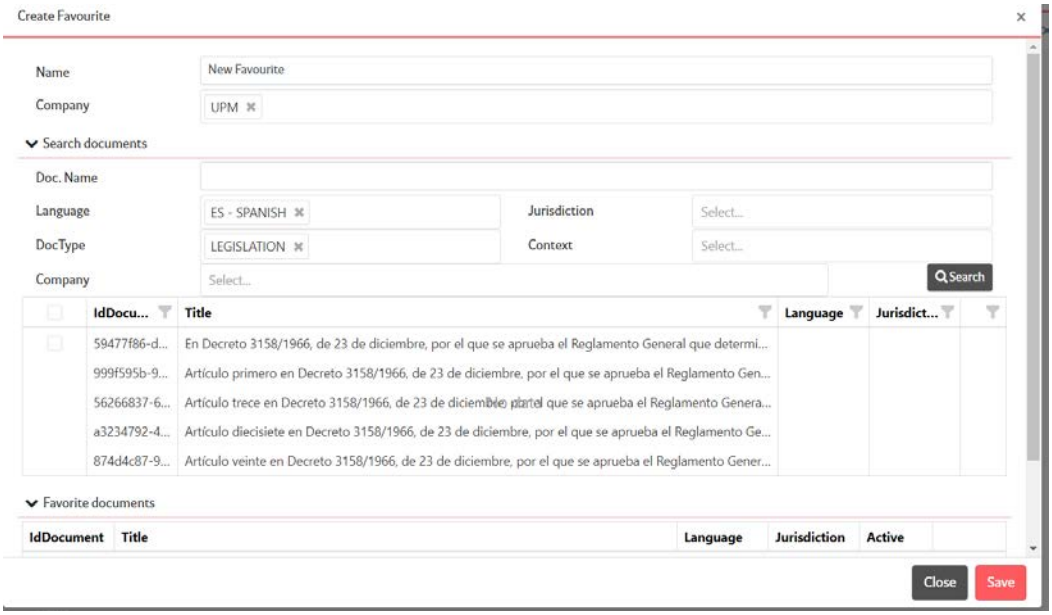


Figure 4. Screenshot User profile. Create Favorite: Selecting documents

Company settings

As a typical user, a lawyer in this case, who works with several clients, the application allows him/her to create several sets of documents for each client or matter. This functionality is implemented combining "Company settings" and defining personal favorite collections into the "User profile".

A Company into the application is a subset of the LKG, a set of documents with a special security. Users can be restricted to work only with some of these sets of documents (typically external users, from this company, but also lawyers who work with some clients should have access to their document collections).

A user (lawyer) who has access to different Company collections can create its own favorite collection with his/her most frequently used documents. These favorite collections can be selected to be used through the document repository selection icon at Question level.

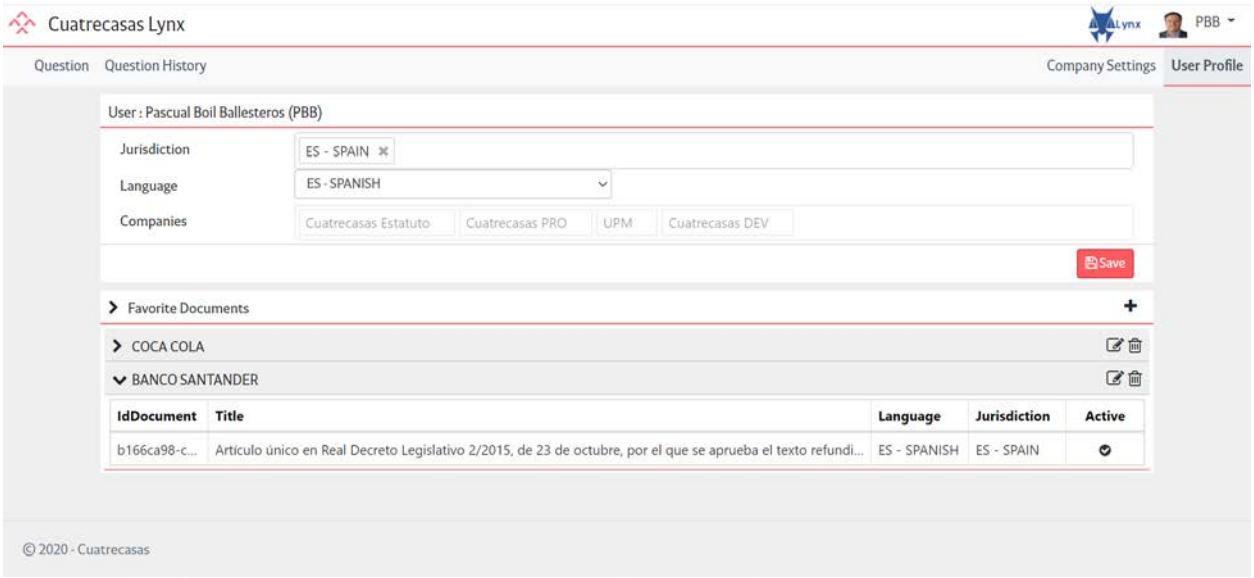


Figure 5. Screenshot User profile Favorites by Company

Users with advanced rights can manage the documents into a Company Collection (today only available for Administrators, but soon available for some specific roles also).

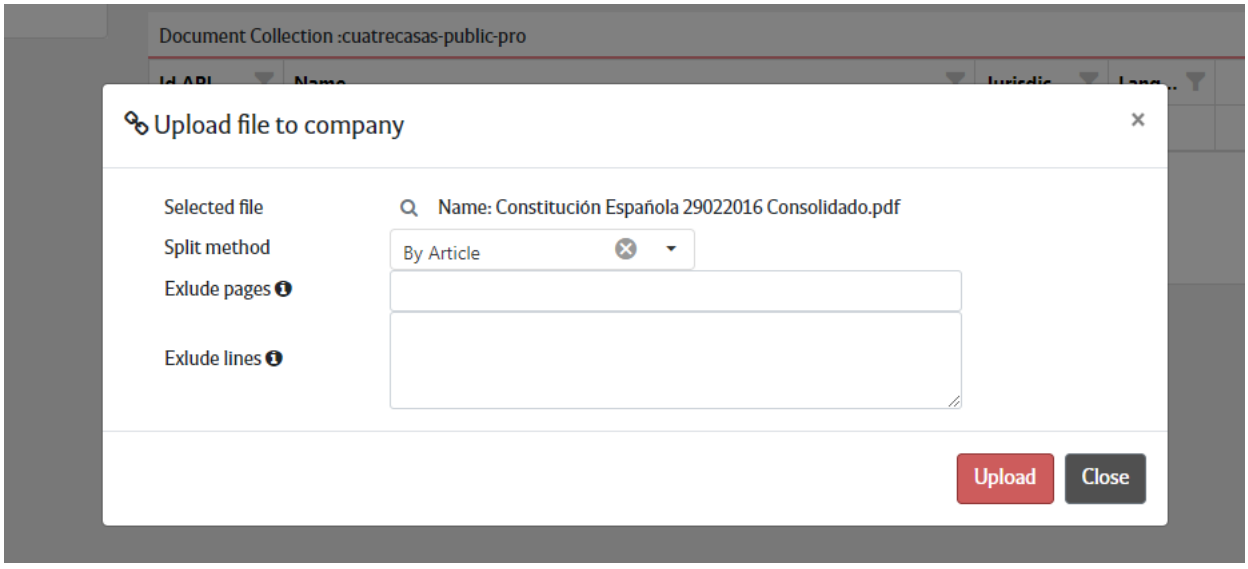


Figure 6. Screenshot Uploading a specific document into a Company collection

When a user uploads a new document into the platform, he or she must choose a split method (split “by Article” is the default option for uploading Laws, and split “by Pages” is the generic for non-laws) (Figure 6). At this point an advanced user can be more accurate specifying the different document division (Figure 7).

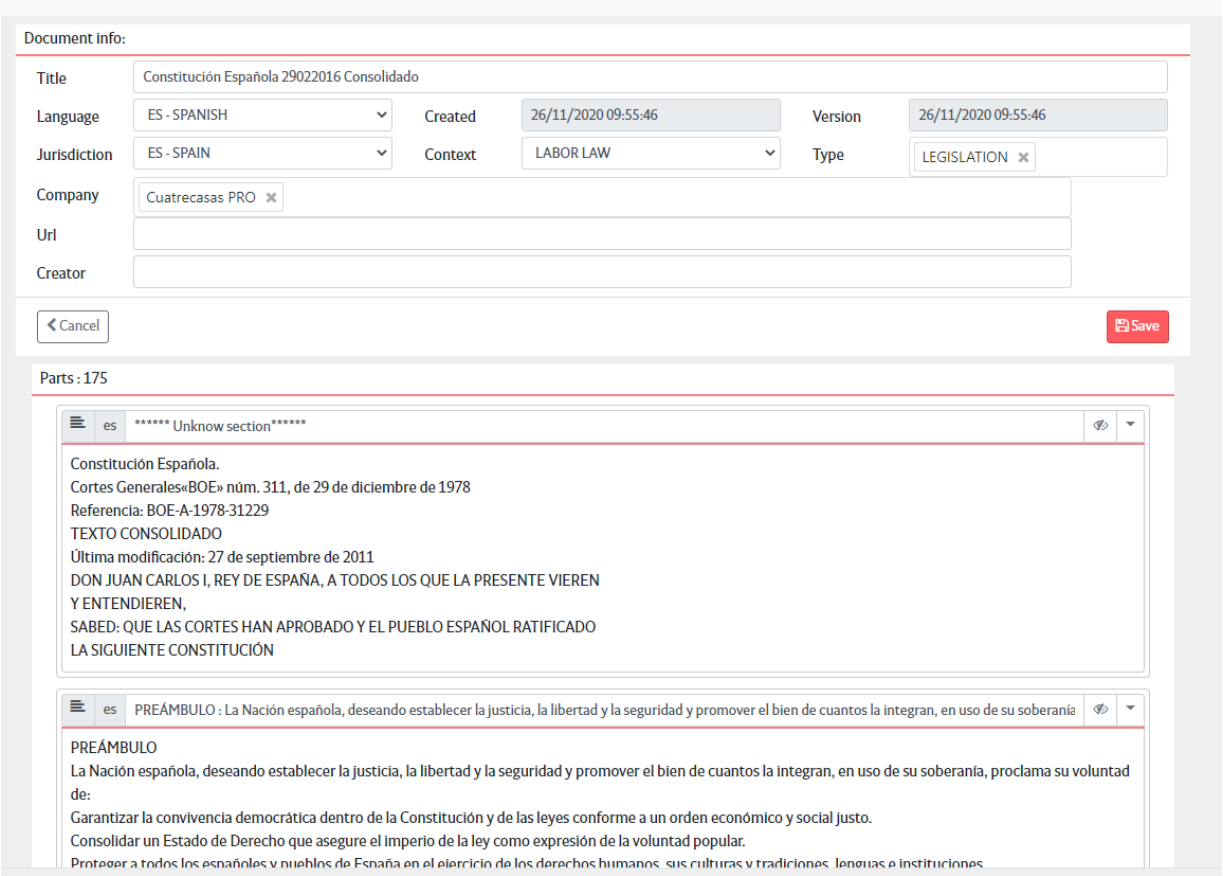


Figure 7. Screenshot upload document splitting by parts

Question

Firstly, a home screen is presented (we are considering to directly go to the "Question" part in the future).

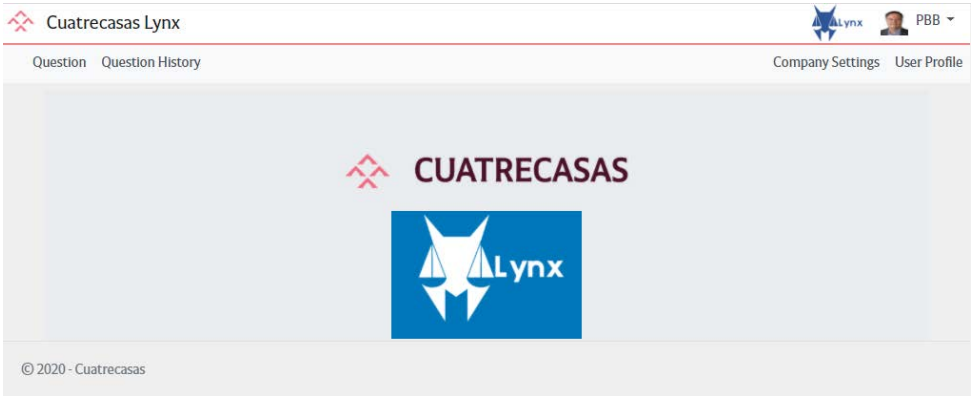


Figure 8. Screenshot Home

Once the "Question" option in the top menu is clicked, the Q&A environment is enabled.

This is the core module of the application where the user can interact with the system writing legal questions related with the document corpus previously selected into the "User profile" and "Company settings" (**REQ06)**.

As a first example (Figure 9) we have entered a question in English, and we have chosen to have a translation "English to Spanish". ("English" is selected as a Question Language and "Spanish" is the user Language Default specified in her user profile).

The question result screen is splitted into jurisdictions. These jurisdictions are defined as part of the "User profile" and "Company settings" configuration (**REQ02**). If documents from different jurisdictions are contained in the corpus, then one tab by jurisdictions will appear.

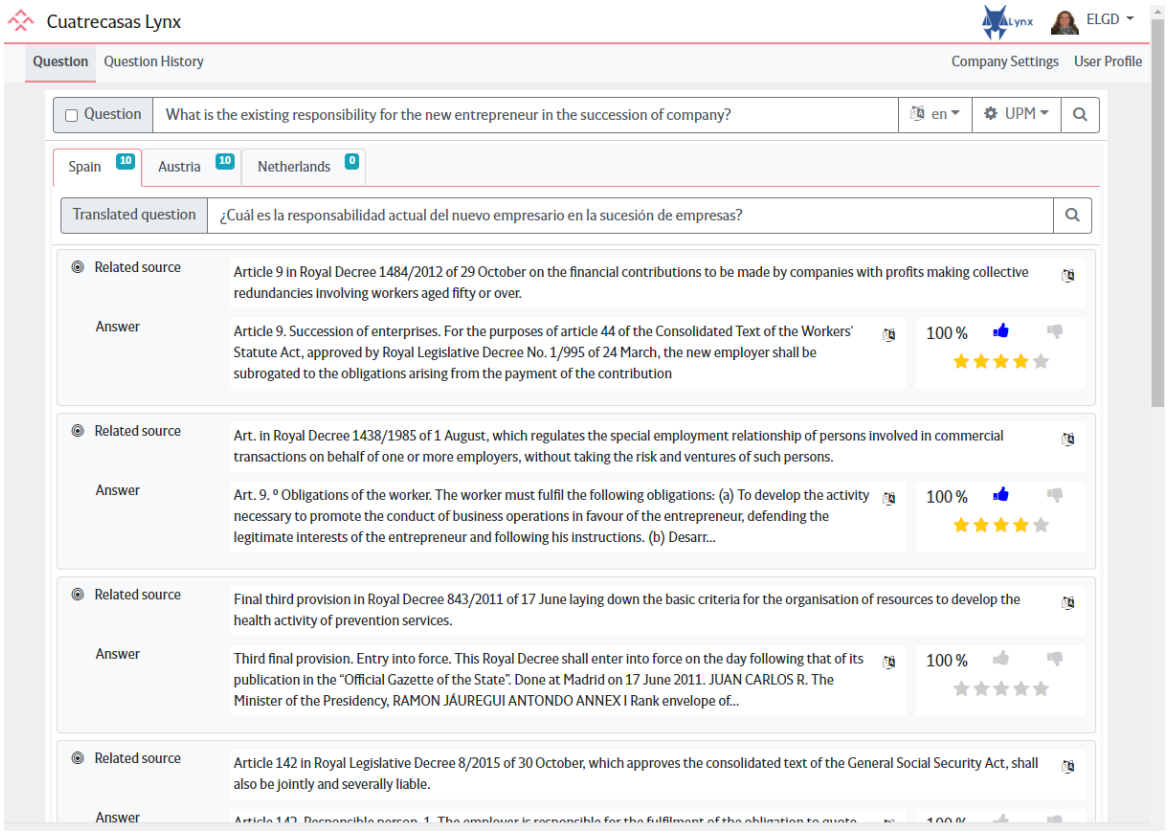


Figure 9. Screenshot Question. Showing results by jurisdiction

Once the question is launched, using the magnifying glasses "Execute search" action, the system is sending all parameters to the SEARCH Service (and also QAdoc) and receiving results that are being parsed and showed to provide additional functionality to the user.

The retrieved results from QADoc service are individual paragraphs from the document complemented (where possible) with a most concrete answer. First part is showed to the user as "Paragraph" and the second is presented today as "Answer" (**REQ08**). The system is also giving and additional information: a number that shows the level of accuracy/trust of the answer. Answers are ranked according to these associated numbers, which are shown to the user to help in the validation and feedback process.

The different results (paragraphs and answers inside) are grouped by "Document" and "Article" as main parts of the document structure.

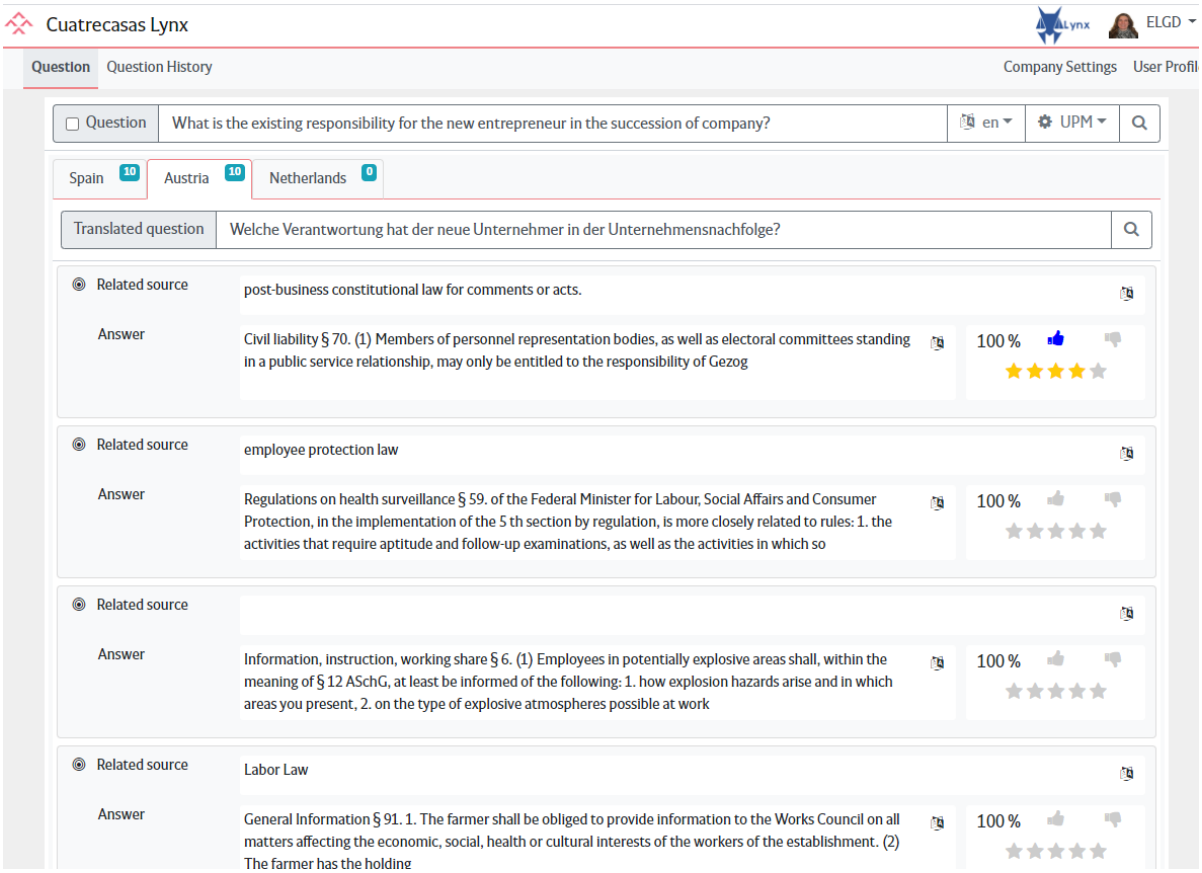


Figure 10. Screenshot Question. Showing results by next jurisdiction

At result or paragraph level, we have also implemented two functionalities:

1. The “Translation” option (through the “world icon”) that allows the user to visualize the translated paragraph text (currently limited to “Spanish”) as shown in Figure 11 (**REQ10**)
2. The “Rate the results” functionality, initially defined simply as a “like” and “dislike”, but today implemented as a “Stars rating” functionality to provide most detailed feedback about the quality level on the answers (**REQ14**).

Figure 11. Screenshot Question. Translate paragraph functionality

As an advanced functionality, the user has the possibility to modify the "Translated question" to include more accurate words in the language of the jurisdiction (if the user knows that language).

We summarize the main options of the Question Screen in this "quick guide" (Figure 12):

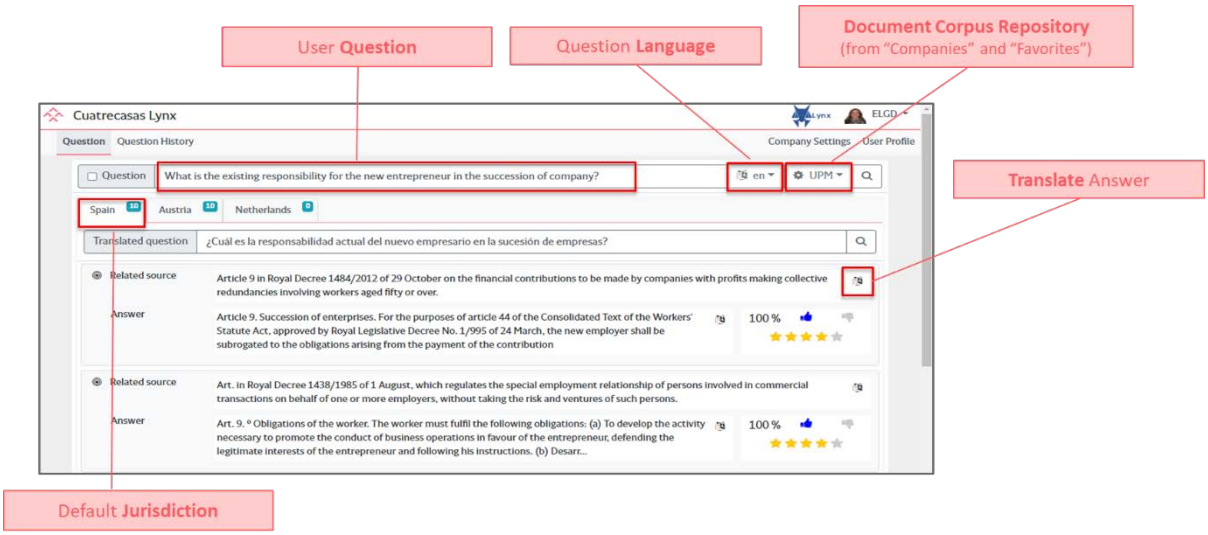


Figure 12. Screenshot with Question Quick Guide

3.5 ACCESS TO DEMO ENVIRONMENT

To evaluate and demonstrate the Cuatrecasas Lynx Application, we have created a demo environment with external access (out of the Cuatrecasas network) restricted by IP address domain to the Lynx Consortium members, only available for a limited time period due to security reasons:

https://labourquery.cuatrecasas.com

- User: CuatrecasasLynx1
- Password: $CGPLynx001

4 PILOT ARCHITECTURE AND TECHNICAL SOLUTION

Figure 13 represents the complete picture of the pilot architecture, with detailed interaction and data flow. The resto of the section details the main components of the architecture proposed for the Cuatrecasas pilot: **frontend, backend, database and lynx API.**

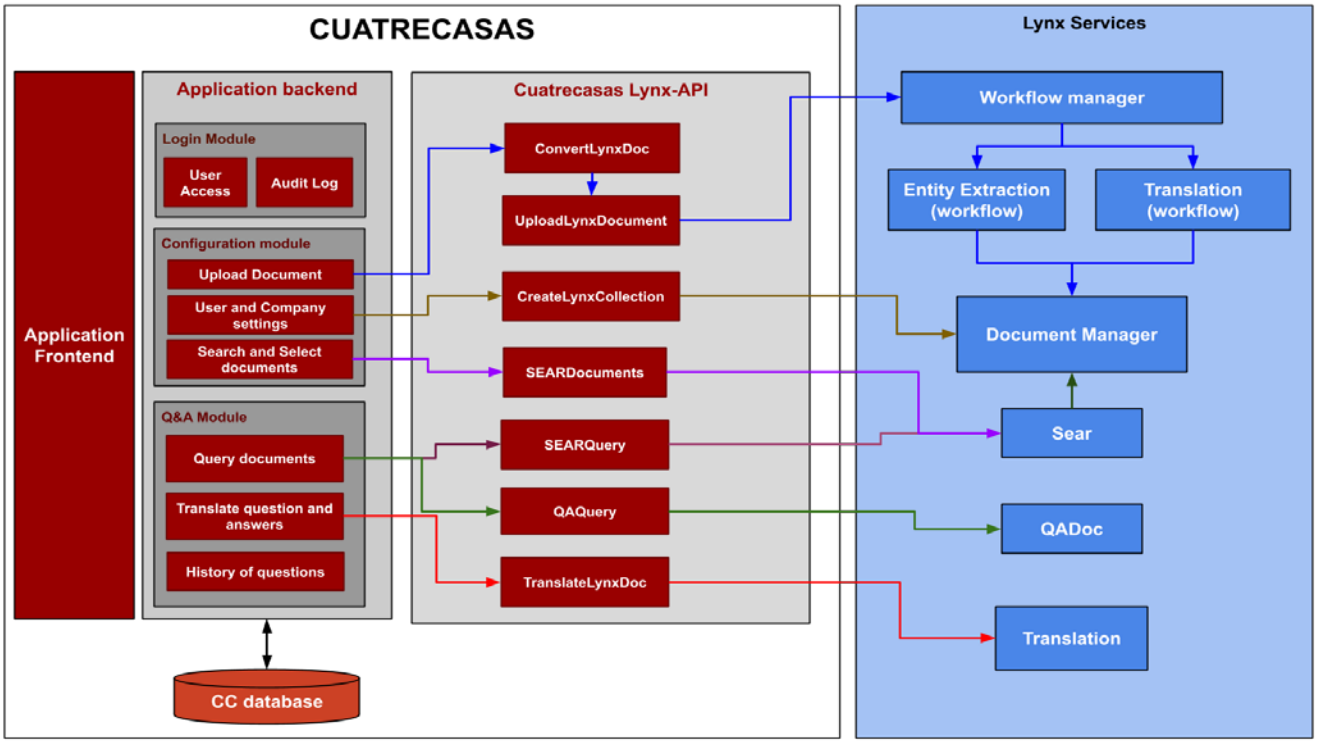


Figure 13. Complete (low level) components schema with services and data flow

4.1 TECHNOLOGICAL SCENARIO

The web server is hosted in a shared web server. This server is hosted in our development environment [SRVWEBINTBDEV] in Azure.

The web server used is Microsoft IIS 7.

The operating system version is Windows Server 2016.

Tools and technologies used for the development:

- ASP.NET C# Core 2.2 as a main programming language
- MS Visual Studio 2017 as developing environment
- MS TFS (Team Foundation) for versioning and code-organization
- Bootstrap 4.0 (*) as a responsive web design (RWD) framework
- JQuery 3.3.1 as a JavaScript library
- DevExtreme 18.2.7 JavaScript UI controls library (part of DevExpress)

Security has been one of the main tasks in the development of the pilot to prevent hacking activities (e.g., obfuscation (masking) of identifiers in all the URLs), and the most common vulnerabilities of web applications such as XSS or SQL-INJECTION.

The application uses two **User Authentication** methods: (i) for internal users, SSO based on ADFS; and (ii) for external users, username and password.

4.2 APPLICATION BACKEND

The application backend is divided into application modules. The main function of these application modules is to provide business logic and solve/implement the functional elements that are used through the UI (User Interface) in the Front-end side. The application functional elements identified are: (grouped by functional module)

- **Login module** [LOGIN]
  - USER ACCESS and Roles
  - AUDIT LOG

- **Configuration module** [CONFIG/DEFAULTS]
  - USER AND COMPANY SETTINGS
  - UPLOAD Document
  - SEARCH and SELECT Documents

- **Question Answering module** [Q&A]
  - QUERY into Documents
  - TRANSLATE Question and Answers
  - HISTORY of Questions

4.2.1 Login module

The functionalities of the login module are:

USER ACCESS and Roles
Users and logical security (document access rights and functionalities) are fully managed by the Cuatrecasas application.

In the application, different roles are defined:

- **End-users**: End users can access the collections based on the company or companies they belong to, with read and write access rights, and with read access to the public collection.
  When a user is created, the user has to be attached to the company or companies it belongs. Each company has their own collections. By default, any user have access the public collection.
- **Cuatrecasas lawyers**: they can access the collections of their clients with read and write access rights, and the public collection with the same access rights.
- **Cuatrecasas Knowledge team**: the members of this team can access all collections available and the public one with full access rights.
- **Administrators**: IT people who can create new collections for new companies, create users, and have full access rights to all documents. In the future this role can be shared with or partially delegated to knowledge lawyers (or some advanced users).

AUDIT LOG
This internal functionality allows the application to store user logging information (who and when) and store some internal usage trace to allow usage statistics and audit requirements.

4.2.2 Configuration module

The functionalities of the configuration module are:

USER AND COMPANY SETTINGS

This functional block is one of the most relevant in the current version of the application. Through this module any user can define his/her own environment and preconfigure the more frequently used repositories.

The main functionalities are listed in the following:

- Define default values for Language and Jurisdiction (to show results in the Question Answering Module)
- Create a personal subset of documents (**Favorite**)
- Create different corpus (subset of documents) for different **Companies** (Cuatrecasas clients or specific matters)

A user could use this option to create several dedicated collections of documents for its own benefit.

UPLOAD Document

This functionality allows the end user to upload PDF and MS Word documents into the Lynx platform (Document Manager through the **Workflow Manager**).

Upload Documents to the Cuatrecasas repository is an advanced functionality only available for Administrators and Knowledge Team roles.

All documents uploaded must belong to a Collection. We decided to have private and public collections, but users are only allowed to upload their own private collections. The public collection is maintained by Cuatrecasas. Some functional conversations and examples about the LKG and DM Collections requirements for this use case is described in Annex 2.

When users upload the Word or PDF document, they must decide which collection the document belongs to and which type of harvester he wants to use (by article or by page). The harvester selected will convert the document into a LynxDocument, splitting it into different parts and creating metadata for each part.

Before uploading the document, users display on the screen the document divided into different parts. The user can change the properties of each part or even create new parts.

Then, the document is converted into JSON format, LynxDocument, enriched with metadata and sent to the workflow manager to be part of the Lynx Platform, so that it can be accessed and queried.

SEARCH and SELECT Documents

By default, user queries will look for an answer in all collections available to the users. Users can also select sub-sets of documents and create favorite subsets to which their queries are sent.

When users decide to create a favorite subset of documents, they can search and select documents.

They can search across all collections they have available based on their permissions and narrow their search by all metadata available for each document.

Once the results of their search are displayed, they can select documents and save their favorite ones.

On the backend, this search is sent to the **SEAR Service** and the relevant documents are received by the front-end. Favorites for each user are saved in their user profile and can be modified or deleted.

4.2.3 Question answering module

The functionalities of the question answering module are:

QUERY into Documents

Query is the main functionality of this application. It is responsible for evaluating and processing the legal question written in natural language and retrieving all possible answers (paragraphs and parts of paragraphs). For that purpose, our application has to combine 2 different Lynx Services: **SEAR Service** and **QADoc Service**.

In this regard, four different question/query types have been identified:

- **Type 1: Training query type.** The question should 'find' the answer in the paragraph, and the answer must have a maximum of 10 words. This kind of questions are typically solved by full-text semantic search engines (like ElasticSearch, which is the one used by the Lynx SEARs). See an example of this type of query in Figure 14.

- **Type 2: Yes/No answer.** The answer to this query is Yes or No. It will be based on finding information very closed in the same sentence.

- **Type 3: Set of paragraphs.** The answer to this query is to be found in multiple paragraphs. The rule is to find (search service) which documents contain the possible answer in order to determine the list of documents that will be queried by QA System.

- **Type 4: Concrete answers (typically on Who? When? Where? How many?).** Direct questions to be solved by a pure Q&A solution, like the QADoc module (in this case exclusively trained in English).

Currently, Type 1 and Type 4 are supported by our Query and retrieve Answers functionality using QADoc and Search Lynx modules.

Question	Paragraph	Answer	Start	End
Who regulates service relations?	a) **Service relations of civil servants shall be regulated by the Statute of Civil Service**, as well as those of personnel in the service of the State, Municipal Corporations and Regional Public Entities, where such relations are governed by administrative or statutory regulations under the protection of a Law.	service relations of civil servants shall be regulated by the Statute of Civil Service	7	100

Figure 14. Sample of question Type 1

TRANSLATE Question and Answers

Language and jurisdictions are managed by the User profile option. Users can decide in which language (from the available list) answers are retrieved, and which jurisdictions are to be queried.

Answers will be translated to the selected language before displaying them in the front-end application. As for the jurisdictions, each one will appear in a different tab.

The translation functionality is provided through the Lynx **Machine Translation Service**.

HISTORY of Questions

The system stores automatically all queries sent in by user, for further analysis and to allow the user to retrieve previous questions.

4.3 DATABASE SERVER (DATABASE)

The Database instance is hosted on a central Database Server (shared by different internal applications). Currently, it is hosted in our development environment [SRVSQL4DEV] in Azure (private cloud).

The DBMS is Microsoft SQL Server 2016.

The operating system version is Windows Server 2016.

The data model represents the main entities identified in our use case (Question, Answer, Company, Jurisdiction, Language, User, …), along with their required fields / attributes. Although not all fields are being used in the current application, the data model is prepared to support new developments (segmentation of companies by sector, organizational information of the company and its work centres, and store relevant information for questions such as the number of employees, whether it has a worker's representative or not, etc. ...). Figure 15 presents the complete representation of the tables and their relations.

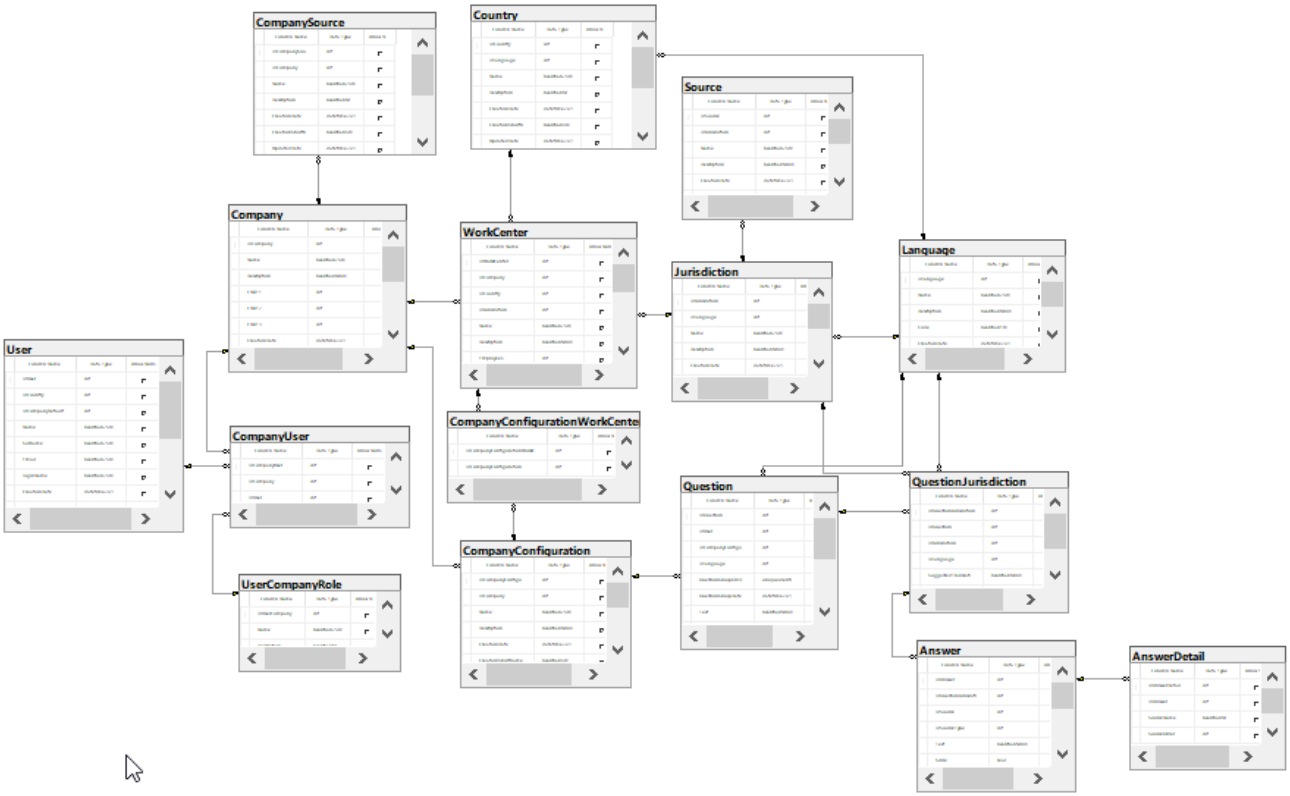


Figure 15. Data model. High level visualization (20-11-19)

4.4 CUATRECASAS-LYNX API

Cuatrecasas has developed a component that works as a macro-service named Cuatrecasas-Lynx API. The component is specialized for this use case but can be used to develop new Cuatrecasas LegalTech applications over the Lynx platform in the future. Cuatrecasas-Lynx API has been designed to make the application layer independent of the specific services behind, of the complexity of the different connectivity methods and of whoever provides the specific services.

This macro-service component manages the calls to the different Lynx services and also contains new methods to improve the communication with them.

This component contains the following methods:

- **ConvertLynxDoc**: converts a document into the Lynx Format.
- **CreateLynxCollection**: creates a new collection into Lynx DCM.
- **UploadLynxDocument**: is connected to the Workflow Manager to upload the document into the DCM, enrich it and translate it if needed.
- **SEARDocuments**: retrieves documents to be selected as favorites and filters them by different properties using SEAR services.
- **SEARQuery**: retrieves a result based on a Question and a document selection using SEAR services.
- **QAQuery**: prepares the result set obtained from the SEAR to be send to the QADoc service to retrieve a better precise answer.
- **TranslateLynxDoc**: using the Lynx Machine Translation Service, translates queries and texts to different languages.

This component can be used by other Cuatrecasas applications to upload documents into the Lynx platform, get more precise answers from a result set and use the Machine Translation service for different languages.

5 PARALLEL ACTIVITIES FOR THE PILOT

The main challenge of this project is to solve legal issues posed in natural language involving labor law and collective agreements. Due to the complexity of the legal language, the several natural languages in which data sources (corpora) are available, and the lack of fine-tuned common services for this domain, some tasks within WP2 and WP3 had to be performed beforehand. In this section, some of these tasks are spelled out.

5.1 IDENTIFY AND STORE THE MOST RELEVANT LAWS RELATED TO LABOR LAW INTO THE LKG

Within the project a considerable effort was devoted to the compilation of laws relevant to the labour area. Identifying the official sources and mechanisms for public access to reliable and updated information in labour law was one of the first challenges we faced in the project. Some high-level explanation about laws hierarchy in Spain is described in Annex 3.

Regarding the Spanish legislation, the collaboration between legal experts in Cuatrecasas and UPM technicians resulted in a first version of the LKG with a large volume of documents in labour law. Labour legislation and royal decrees were uploaded in a first stage, prioritizing consolidated versions published in the Spanish Official State Gazette (BOE). In the next stage, sectorial collective agreements at national and regional levels were also contributed to the LKG. Overall, several thousands of documents related to labour law in Spain became part of the LKG. Finally, legislation and collective agreements from Austria and other countries in Europe were also added to the LKG.

5.2 CREATION OF DATASETS

During the project, Cuatrecasas has also created datasets to train, fine-tune and evaluate other Lynx services. Such produced datasets contain expert knowledge and are high valuable resources for the scientific community.

1. **Question and Answers Dataset.** A collection of questions and answers have been created manually with the effort of Cuatrecasas legal experts. These questions and answers are related to the Spanish Worker's Statute document. For each question, following information items are added:
    a. Section: section of the Spanish Worker's Statute which contains the article of the answer
    b. Article: article of the Statute that contains the answer
    c. Paragraph: paragraph of the article that contains the answer
    d. Answer: sentence of the paragraph in bold which answers the question

    The dataset contains every field (questions, sections, articles, paragraphs and answers) in English and Spanish. The number of questions amounts to 400 questions distributed in different Excel spreadsheets.

    The resulting dataset was used a first step to fine-tune the QADoc model. Moreover, the dataset was also used to evaluate the quality of the pilot and the involved services (as described in Section 5.4).

    The dataset has been publicly published in the Zenodo data portal as a contribution of the Lynx project (https://zenodo.org/record/4256718#.X8TRHmhKiUk), and has been submitted to a conference for its dissemination (Calleja et al., 2020). At different stages of the project, legal experts have manually revised the dataset, which makes it a valuable result of the project (see Annex 5 with some feedback examples as part of the QADoc training).

2. **Real Labour documents dataset** for Machine Translation. A set of official documents (most of them public, other private but anonymized) related to labour domain were collected and provided to train domain specific automatic translation services for the language pair Spanish-English. The set contains documents in Spanish and English languages, as well as bilingual documents with content in both languages (two-columns documents) manually translated by certified translators.

5.3 CREATION OF A LABOUR LAW TERMINOLOGY

Since the beginning of the project, legal terminology was considered a key aspect to provide multilingual expert knowledge. Terminology harvesting and generation efforts for the three Lynx pilots are duly documented in D2.2 and D2.7, including semi-automatic approaches for terminology creation and enrichment provided by Tilde and UPM. In this case, such approaches have been supported by senior lawyers and members of Cuatrecasas Knowledge Team.

In this pilot, the terminology work is specialised in labour law, and consequently, the main working document has been the Spanish Workers' Statute[1]. The work started by automatically identifying the most relevant terms of this document, helped by different term extraction tools: Tilde's services, SketchEngine, TermSuit and TBXTools. For each term, a context of use was also extracted, so it could be used to disambiguate when necessary, that is, to retrieve the correct sense of the term from external resources.

The need for such a disambiguation step lies in the "enriching stage" of the terminology generation process, in which several existing language resources have been queried to retrieve additional linguistic information related to each of the terms previously extracted: translations, synonyms, definitions, hypernyms, hyponyms and other related terms. Some of those existing language resources are of a more general domain, thus containing ambiguous data (such as Wikidata[2], IATE[3] and the KDictionaries[4]), and others are specialised in the legal domain (such as EuroVoc[5], STW[6], Unesco[7], ILO[8] and Thesoz Thesaurus[9]).

The main purpose of retrieving additional data, specifically synonyms or term variants, is to contribute to the query expansion process implemented in the Question and Answering Module (SEAR and/or QADoc services). Also, by retrieving translation, a monolingual plain list of terms was converted into a multilingual terminology to be used for navigation purposes amongst documents in different languages.

The creation of the labour law terminology has been semi-automatically performed and manually reviewed by Cuatrecasas legal experts. Several versions of the terminology, which is represented following the SKOS vocabulary[10], have been accordingly published in Zenodo[11].

The final version of the terminology is also published in PoolParty, because it allows an easy visualization and management (see Figures 16 to 18).

[1] https://www.boe.es/buscar/act.php?id=BOE-A-2015-11430

[2] https://wikidata.org/

[3] https://iate.europa.eu/home

[4] https://api.lexicala.com/

[5] http://eurovoc.europa.eu/

[6] https://zbw.eu/stw/version/latest/thsys/71055/about

[7] http://vocabularies.unesco.org/browser/thesaurus/en/

[8] https://metadata.ilo.org/thesaurus.html

[9] http://lod.gesis.org/thesoz

[10] https://www.w3.org/TR/swbp-skos-core-spec/

[11] https://zenodo.org/record/3843561#.X8UO8apKi3I

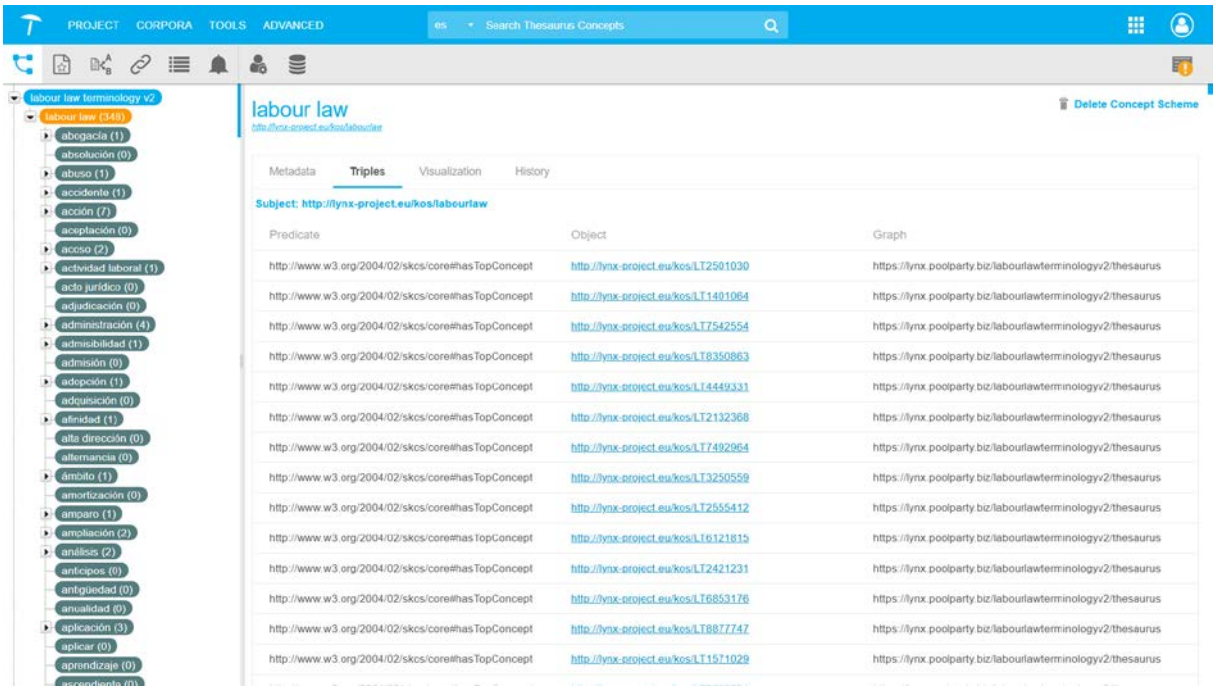


Figure 16. Labour Terminology general view

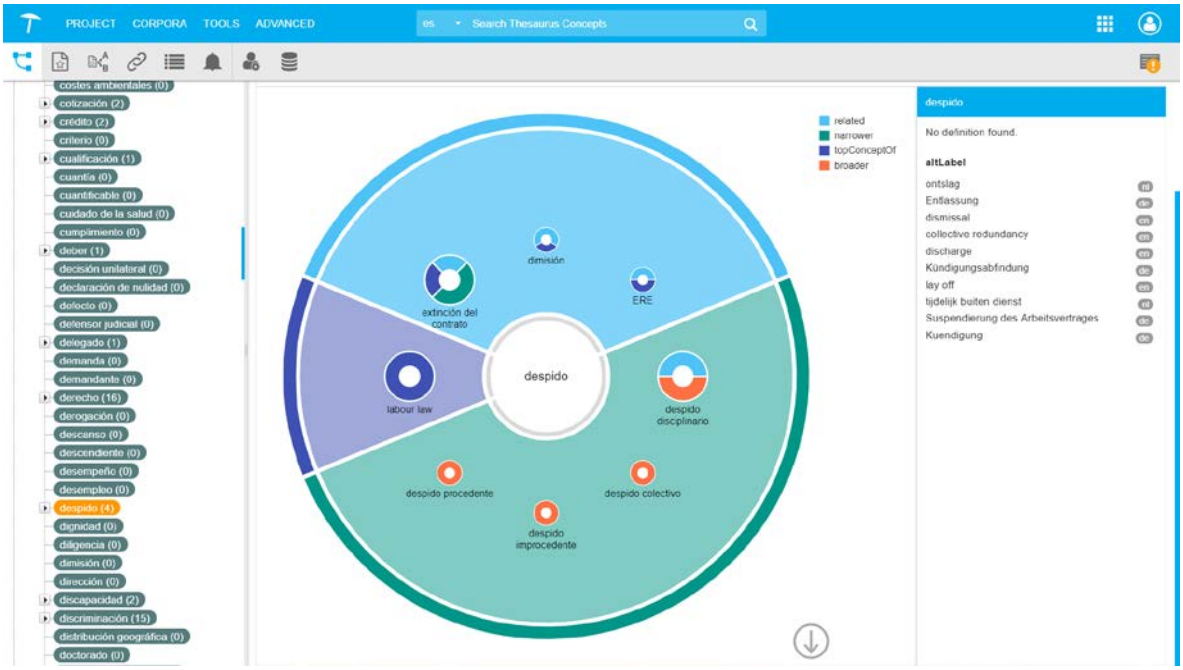


Figure 17. Labour Terminology extract "lay off" graphical view

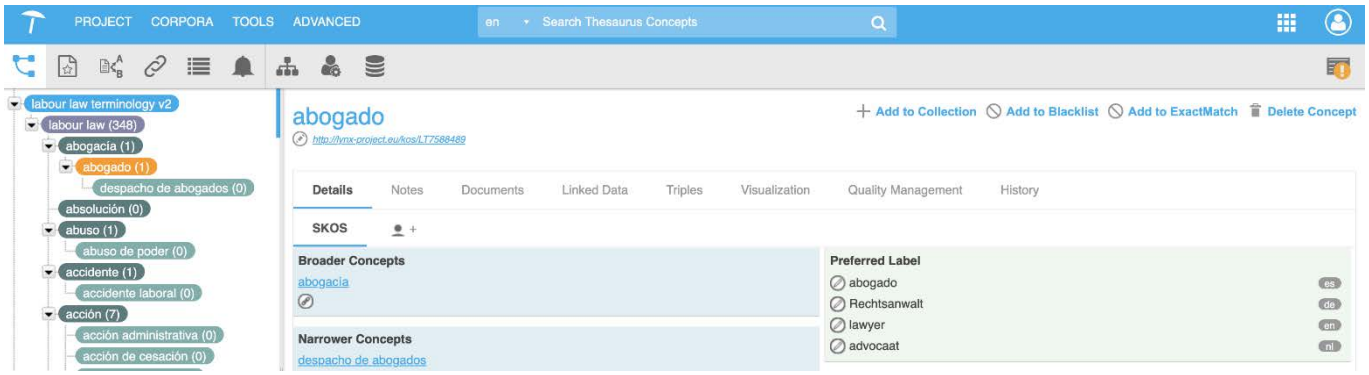


Figure 18. Labour law term entry "abogado" (lawyer) with translations into 4 languages

Currently, SEAR services from UPM and Cybly are evaluating the impact of using synonyms and related terms from the terminology to retrieve documents. Evaluation results will be included in D5.8.

5.4 TIME EXPRESSION ANALYSIS FOR THE LEGAL AND LABOUR LAW CONTEXT

As part of the UPM research on Time Expressions and with the objective to train the TimEx service, an extensive document was created with all relevant time expressions in the Spanish Worker's Statute (in Spanish). This work is reported in D3.8 (Annex 2).

5.5 EVALUATION SCENARIO

As part of the continuous integration tests performed by UPM during the project, UPM has created an evaluation scenario in which part of Cuatrecasas targets are addressed.

This evaluation scenario consists in using 150 of the questions of the Question and Answers dataset (Section 5.2) related to the Spanish Workers' Statute to measure the precision of the UPM's SEAR. The scenario reads each question in Spanish language and sends it to the UPM's SEAR. Then, the SEAR service returns up to ten documents from the specific collection that contains the Spanish Worker' Statute segmented by articles. Each article is a LynxDocument. The evaluation measures if the target article (the answer) is included in the set of candidate documents that SEAR retrieves. UPM's SEAR is prepared to receive a full question formulated in natural language and internally performs a Query Expansion mechanisms to improve the identification of the most suitable answer.

The evaluation has been addressed by three experiments in relation to the Query Expansion mechanisms that UPM uses in their SEAR service. The first experiment consisted in using the question in its original form (a natural language question). In the second one, the question is enriched repeating those terms that are in the question and in the extracted terminology (relevant terms of the domain, Section 5.3). Finally, in the last experiment, the question is enriched as in the previous experiment, and also with the terms identified by the Rake library (https://pypi.org/project/rake-nltk/). The purpose of this library is to identify the main terms (nominal chunks) that appear in the question to reflect their importance in the scoring process.

The result of each question is a list of up to ten documents which are classified into four groups: perfect, correct, others and not found. "Perfect" means that the target document appears as the first result in the list (the most important one); "correct" means that it is contained in the firsts four results; "others" means that the target document is contained in the list, but not as part of the first four results; finally, if it does not appear in the list, it is classified as "not found". The experiment results are presented in Figure 19. Results expressed with the dot are percentages and the results in parenthesis are absolute numbers of documents.

	Perfect	Correct	Others	Not Found
Experiment 1. Q	.66 (99)	.18 (27)	.11 (17)	.04 (6)
Experiment 2. Q+ET	.66 (99)	.18 (28)	.1 (15)	.04 (7)
Experiment 3. Q+ET+ATE	.68 (103)	.16 (24)	.1 (15)	.04 (7)

Figure 19. Experiment results for 3 Query Expansion methods

The results show that even with the simplest query in Experiment 1, a high performance in "perfect" and "correct" answers is reached. Surprisingly, the use of the extracted terminology in Experiment 2 has not affected the overall results; only one question has passed from the category "others" to "correct". Future experiments will analyze the use of the synonyms reflected in the terminology to enrich the query. However, the combination of the extracted terminology and the automatic term extraction over the query has reached the highest performance score with 103 documents (68%) classified as "perfect".

6 OUTLOOK

6.1 CURRENT EXPERIMENTS

Currently our focus is on evaluating and improving the quality of the results. In this regard, the activities we are engaged in are:

- Working with UPM and Cybly (former openlaws) to improve search results, experimenting with different strategies on query expansion techniques over SEAR service interaction.
- Enriching Lynx document metadata with annotations about legal concepts and complementary structure of the document.
- Working on the evaluation of the QAdoc service and its combination with the Translation service to assess if it is possible to avoid training the former for each new language.

The results of these experiments shall be reported in D5.8.

6.2 NEXT STEPS

The paragraphs below describe the next steps to be done in order to exploit the pilot after the project end. We have grouped them considering the time horizon.

Short- and medium-term actions

1. **Improve Search Service filtering to work efficiently with a higher number of documents**. As with any other information retrieval problem, increasing the base of documents makes more difficult to keep good precision figures. This problem can be mitigated with better filtering.
2. **Include in the knowledge base a predefined set of questions and answers provided by Cuatrecasas legal experts.** Cuatrecasas already has a set of frequent question-answer pairs, elaborated by expert lawyers. This expert knowledge should be injected in the system. The pre-defined answers will work as Favorite or Recommendation from the expert, and they would be presented in the form of a featured or “sponsored” result.
3. **Specialize the existing algorithms to tackle different question types.** Whereas current algorithms do not distinguish the type of question, future developments may explore specialized techniques.
4. **Explore and improve new Query Expansion techniques,** such as including Semantic Similarity in question and possible answers.
5. **Improve the QADoc module with more questions**. A very large training set of question-answers (between 500 and 1.000) may boost the quality of the results.
6. **Create specific external interface for customers**, including previous workshop sessions with potentially interested customers to show the application and brainstorm together co-creating their more valuable functionalities (additional security measures would need to be implemented, ethical hacking test). An important part is to analyze in detail the market potential, which can be very different depending on the level of quality of the responses we get. What we initially designed as a tool for our clients could be a more general or basic independent product for SME's or even smaller law firms.

Long-term actions

1. **Include related jurisprudence identification based on each answer**. Search results may also include jurisprudence, obtained in real-time.
2. **Include additional supported languages and jurisdictions (France, Italy, Portugal, UK, …).** This task is related to the improvement and adaptation of some Lynx services (such as machine translation or the query expansion in search) for new languages.
3. **Introduce other disciplines of law (Intellectual Property and GDPR, Tax, …)**

7 REFERENCES

1. Calleja, P., Martín-Chozas, P., Montiel-Ponsoda, E., Rodríguez-Doncel, R., Gómez, E. and Boil, P. (2020). Bilingual Dataset for Information Retrieval and Question Answering over the Spanish Workers Statute. Submitted to the 9th AICOL Workshop (Artificial Intelligence and Complexity of the Law) within JURIX 2020.

ANNEX 1 – SAMPLES OF REAL QUESTIONNAIRES

Cuatrecasas has hundreds of real examples with questions regarding labour law, usually relating to international M&A operations, some led by us, but most led by other firms.

We sometimes have a long questions list, as in this first example: a big US technology company would like to open a new R&D centre in Europe. They create a list with approximately 200 questions to evaluate 5 countries (the UK, Spain, Italy, Germany and the Netherlands) and 20 related exclusively to conditions to contract students as interns/trainees.

In the second example (below), there is a short list of questions, but they have to be evaluated against more than 50 countries (international industrial company that is evaluating the acquisition of another big company with delegations and factories around the world).

Confidential information
Draft subject to review
02/06/2015

Country[1]	Employees	Representatives	Obligations on the Merger decision	Date	Obligations on change of employer[2]	Date
Austria	56	0	ACI has to inform employees on the Merger. Employees cannot block the operation. Infringement could entail fines.	Before the shareholders' agreement on the merger.	ACI has to inform affected employees on the transfer. Employees cannot block the operation. Infringement could entail fines.	Usually one month before the transfer takes place.
Belgium	277	1	ACI has to inform and consult with the workers' representative. Employees cannot block the operation. Infringement could entail fines, damages, and could constitute a criminal offence.	Before the shareholders' agreement on the merger, consultation should last 1 week.	ACI has to inform and consult with the workers' representative. In practice, there is only one information and consultation process with the works council, relating to both the merger and its consequences for the employees. From an HR perspective, ACI could inform the employees. Employees cannot block the operation. Infringement could entail fines, damages, and could constitute a criminal offence.	Before the shareholders' agreement on the merger, consultation should last 1 week.

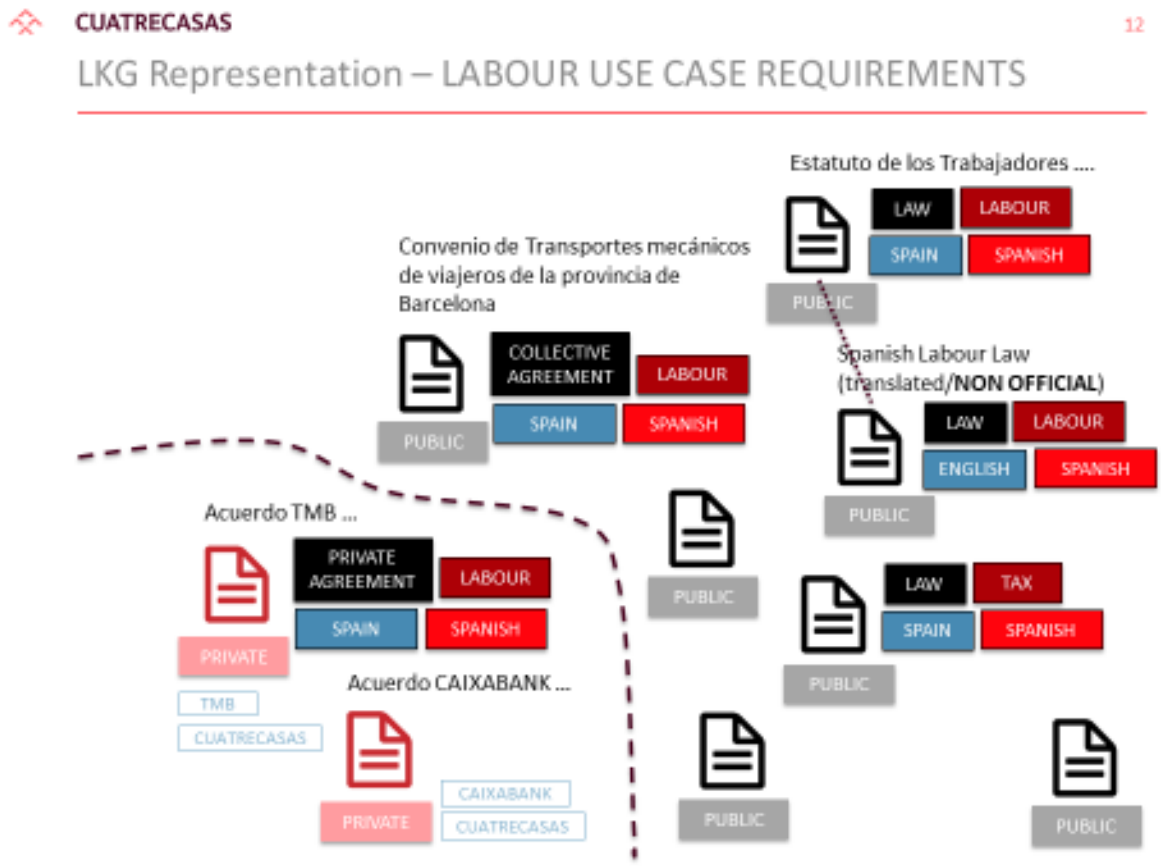
ANNEX 2 – LKG IN THE LABOUR USE CASE

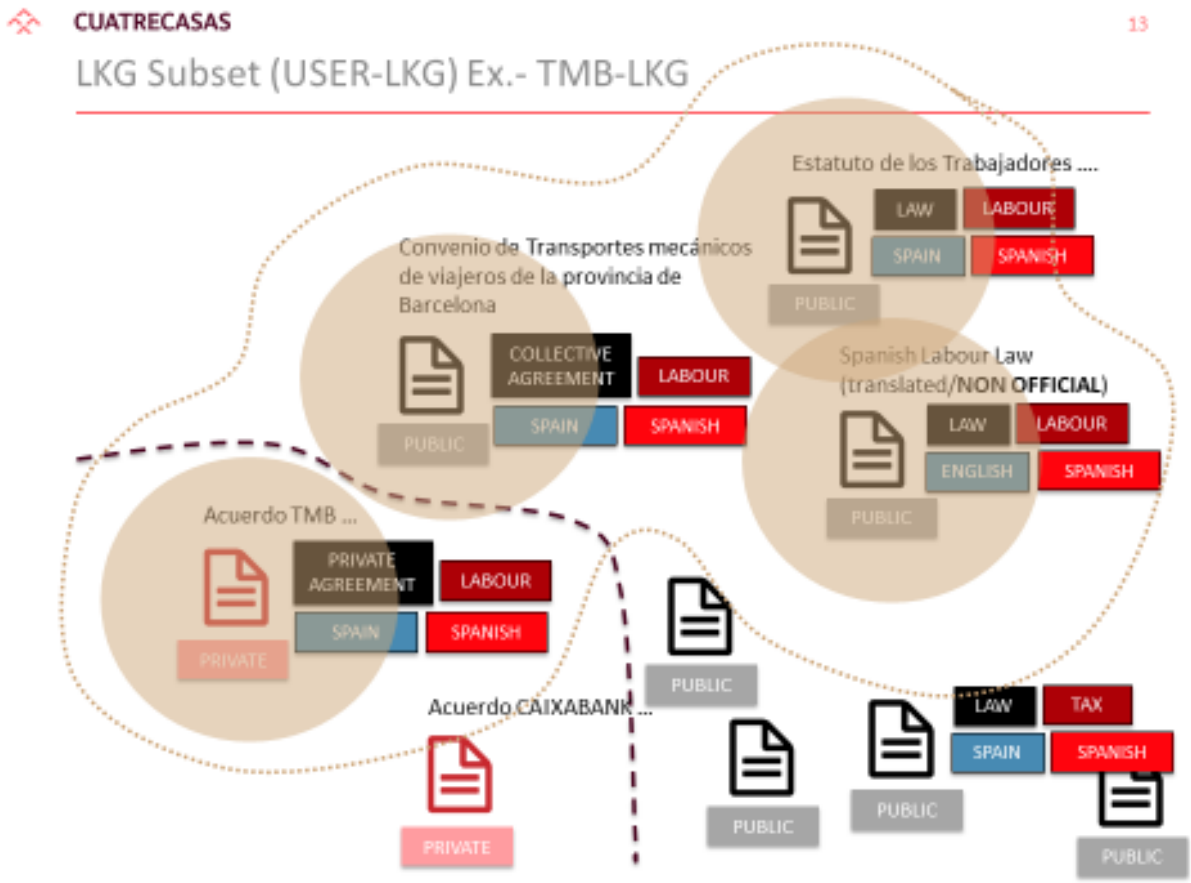
In this Annex, we try to conceive how the LKG from the labour law pilot perspective should evolve. We envision an LKG where we could find the following:

- Documents/content of different types (e.g., laws, contracts, labour sector agreement, and company agreement) for different use cases in the legal and regulatory/compliance ecosystem, as well as for different business contexts (e.g., labour, geothermal and real state). We define this as metadata classification.

- Documents with different levels of security. We assume that LKG will have plenty of public documents and legal resources, as well as private ones. In the current implementation, access to collections is controlled in a per-company basis, but more diverse access control alternatives might be explored, possibly fine-grained.

In the two diagrams below, we illustrate how we could pass from the general LKG (which shows the part related to the labour domain) to one specific application execution logical subset of the LKG for the use case of one specific company (TMB).




CUATRECASAS
13
LKG Subset (USER-LKG) Ex.- TMB-LKG
Estatuto de los Trabajadores ....
LAW
LABOUR
SPAIN
SPANISH
PUBLIC
Convenio de Transportes mecánicos de viajeros de la provincia de Barcelona
COLLECTIVE AGREEMENT
LABOUR
SPAIN
SPANISH
PUBLIC
Spanish Labour Law (translated/NON OFFICIAL)
LAW
LABOUR
ENGLISH
SPANISH
PUBLIC
Acuerdo TMB ...
PRIVATE AGREEMENT
LABOUR
SPAIN
SPANISH
PRIVATE
Acuerdo CAIXABANK ...
PRIVATE
PUBLIC
PUBLIC
PUBLIC
LAW
TAX
SPAIN
SPANISH
PUBLIC

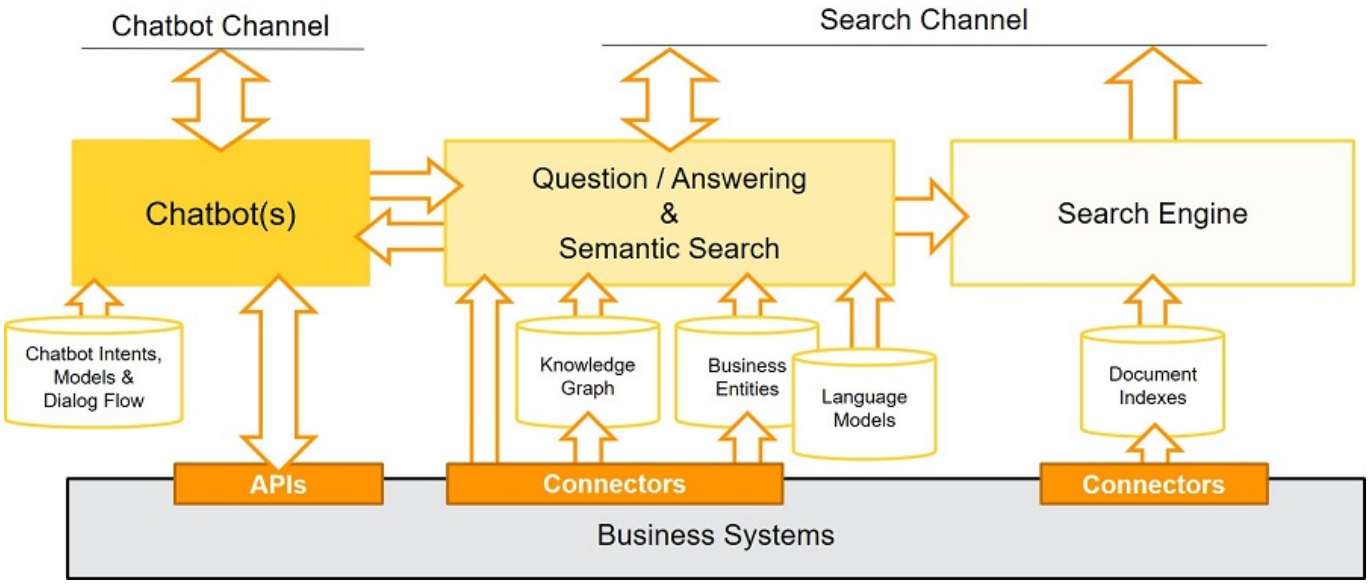
ANNEX 3 – CHATBOTS, QA AND SEMANTIC SEARCH WORKING TOGETHER

About Legal Chatbots

Chatbots are a hot topic right now in many industries, including in law. In recent years, there has been a lot of hype around chatbots such as DoNotPay (https://donotpay.com), helping people get legal help without a lawyer or even talking to a real person. The field of legal chatbots has since expanded and now encompasses a diverse group of bots that use different methods and have different target audiences.

Chatbots, Question Answering (QA), and Semantic Search – How Do They Work Together?

Chatbots handle deep dialogs and specific domains while QA systems handle broad domains of knowledge. However, chatbots and QA systems can be complementary depending on the interface where the user starts to look for information (a search box first or a chatbot interface first). Semantic search and search engines can be used as fallbacks. Based on costs, the depth of knowledge, and other potential criteria required by your organization, an assessment could help you select one or a combination of these solutions.



The architecture diagram above showcases how chatbots, QA, and search interact with the business system and each other to create a fully integrated, intelligent knowledge system within the enterprise.

- **Chatbots** provide deep dialogs to help perform specific tasks.
- **QA systems** interface with business systems and knowledge graphs to answer questions.
- **Semantic search** understands what you are searching for and returns highly-targeted documents or records.
- **Search engines** find documents that best match the list of words and tokens from the user.

Using AI and NLP techniques to merge content sources and create knowledge graphs, we can then leverage chatbots, QA, and search to deliver holistic knowledge and understanding of the enterprise to the user. This is where we think the industry is heading to.

(*) See original complete document in https://www.searchtechnologies.com/blog/search-chatbot-question-answering

ANNEX 4 – USER VALIDATION AND TRAINING FEEDBACK

With the final version of the application, we have been able to contribute to our legal experts' test and validation activities. Below is one example of detailed feedback, which explains not only what the expected results should be, but also indicates key parts of the document that are not currently being considered by the QADoc Service.

In the Figure below, we provide an example of feedback that could be used to train and improve the QADoc service.

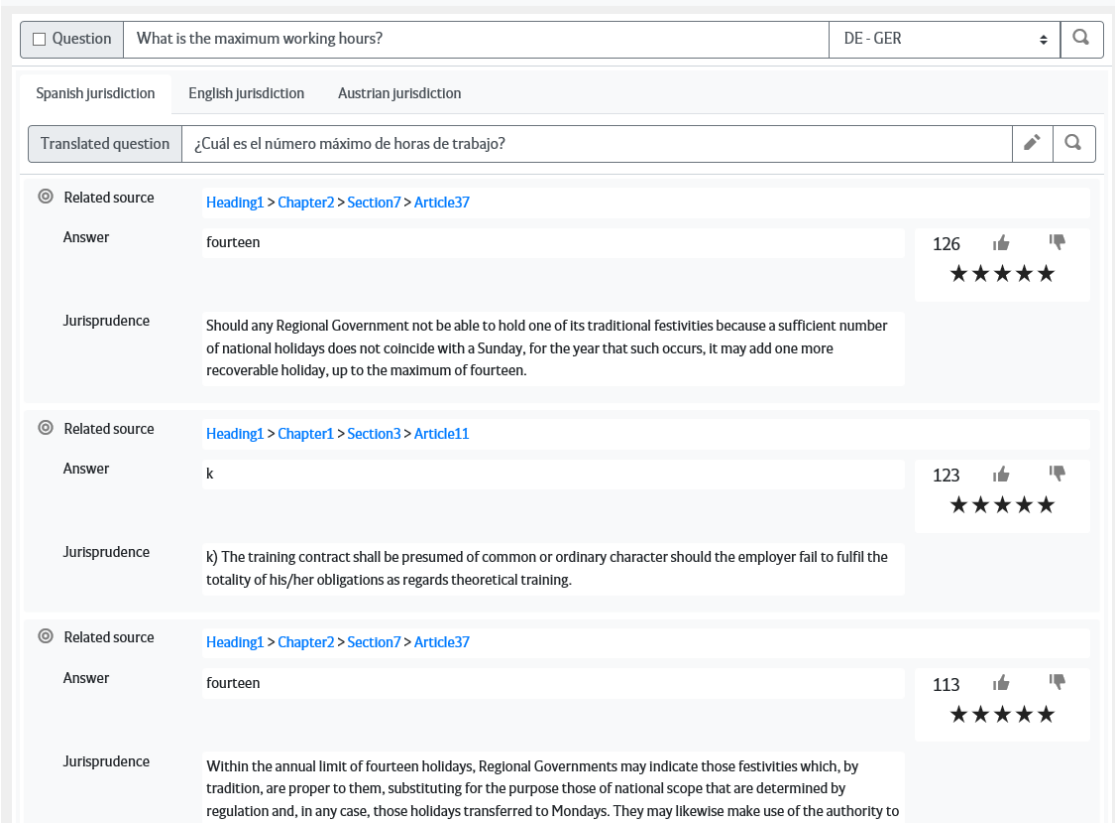


Figure 20. Screenshot 16-10-19 (part of Test and Feedback for QADoc)

We are collaborating with SWC (and Openlaws/Cybly) to improve accuracy. Based on the content of the law, we are also helping them to (i) identify the ideal results to be returned, and (ii) determine the best way to manage and present these results.

We provided detailed feedback in PowerPoint format, not only indicating the right answers, but also providing information on how to prioritize answers or discard results based on the context we can deduct from the documents' structure (articles and sections). The Figure below shows how the descriptions of the articles are not related to the main concept of the question.

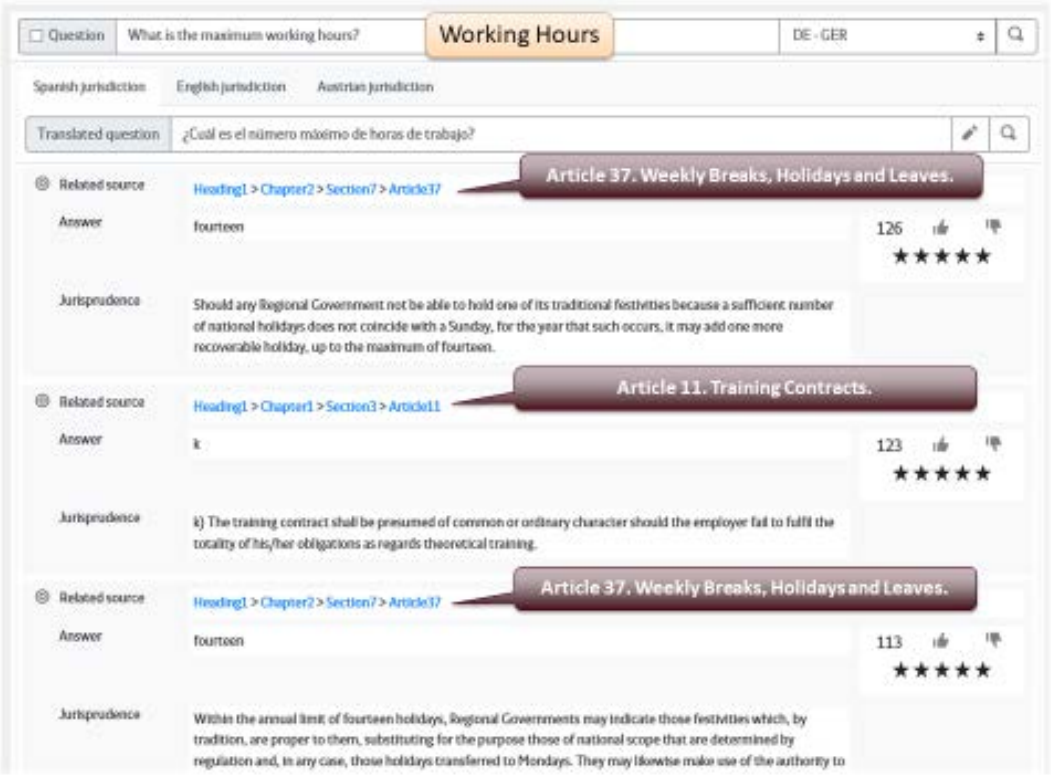


Figure 21. Commented screenshot (part of theTest and Feedback for QADoc)

In following slide, we also show potentially good answers based on the information provided by experts. In this example, we explain the key information in the texts of Spanish labour law that relates directly to the concepts behind the user's question. In this exercise, (see Figure below) for each paragraph, we indicate the different types of results that the system will provide: precise answer (in bold red) and part of the sentence that the use could consider as relating directly to the question (in bold) to be highlighted by the front-end application.

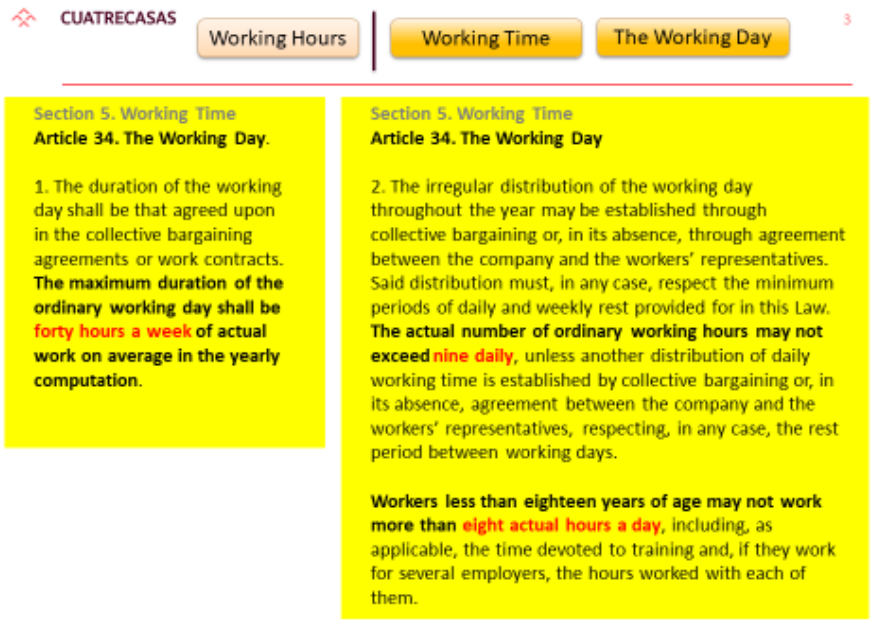


Figure 22. Explanation slide (part of the Test and Feedback for QADoc)