Proceedings of the 12th Conference on Language Resources and Evaluation (LREC 2020), pages 2332–2340
Marseille, 11–16 May 2020

Orchestrating NLP Services for the Legal Domain

Julián Moreno-Schneider[1], Georg Rehm[1], Elena Montiel-Ponsoda[2], Víctor Rodríguez-Doncel[2], Artem Revenko[3], Sotirios Karampatakis[3], Maria Khvalchik[3], Christian Sageder[4], Jorge Gracia[5], Filippo Maganza[6]

[1] DFKI GmbH, Germany – [2] Universidad Politécnica de Madrid, Spain – [3] Semantic Web Company GmbH, Austria
[4] Openlaws GmbH, Austria – [5] Universidad de Zaragoza, Spain – [6] Alpenite, Italy

Corresponding author: Julián Moreno Schneider – julian.moreno_schneider@dfki.de

Abstract

Legal technology is currently receiving a lot of attention from various angles. In this contribution we describe the main technical components of a system that is currently under development in the European innovation project Lynx, which includes partners from industry and research. The key contribution of this paper is a workflow manager that enables the flexible orchestration of workflows based on a portfolio of Natural Language Processing and Content Curation services as well as a Multilingual Legal Knowledge Graph that contains semantic information and meaningful references to legal documents. We also describe different use cases with which we experiment and develop prototypical solutions.

Keywords: Text Analytics, Tools, Systems, Applications, Knowledge Discovery/Representation

1. Introduction

We present a methodology and tooling to handle a set of various Natural Legal Language Processing and Document Curation services currently under development in the EU project Lynx[1]. First, the platform is acquiring data and documents related to compliance from jurisdictions in different languages with a focus on English, Spanish, German, and Dutch along with terminologies, dictionaries and other language resources. Based on this collection of structured data and unstructured documents we create a multilingual Legal Knowledge Graph (LKG), represented as Linked Data. Second, a set of flexible language processing services is developed to analyse and process the data and documents to integrate them into the LKG. Semantic processing components annotate, structure, and interlink the LKG contents.
The following research challenges arose during the project:

(i) How to efficiently orchestrate a set of NLP services? How to guarantee their interchangeability?

(ii) How to efficiently extract information and store documents along with the extracted information?

(iii) In which business scenarios are legal NLP services able to generate actual added value?

The remainder of this article is structured as follows. Section 2 describes different use cases, addressing challenge (iii), while Section 3 focuses upon the LKG used in the prototype applications, addressing challenge (ii). The infrastructure, semantic services and their orchestration through the content and document curation workflow manager, is described in Section 4, addressing challenge (i). After a brief review of related work (Section 5) we summarise the paper and describe future work (Section 6).

[1] http://lynx-project.eu

2. Use Cases

The following three briefly sketched use cases illustrate the development work in the project. The pilot initiatives described next are still not operational, but the current efforts will make them operational in the summer of 2020.

The objective of the *Contract Analysis* use case is to enhance regulatory compliance and obligations through automation, reducing costs, corporate risks and personal risks. Currently, companies have to manage large amounts of heterogeneous contracts, which is, typically, a time-consuming manual process. SMEs do not use management systems to help them in identifying the core data and main actions enforced by a contract. Usually, only a minimal amount of information is kept in a spreadsheet (title, parties, date of signature), which is insufficient to effectively manage contracts and keep track of the actions that need to be taken by the company (such as renewal or amendments). To ensure compliance, improve governance and mitigate risks, companies need to rely on systems that support them in digitizing contracts, identifying the core data, providing an overview of the main content, pointing to the relevant legislation in force, and sending out notifications in case actions need to be taken. Some work has already been done in this sense (Chalkidis and Androutsopoulos, 2017). The prototype will extract information from contracts and subsequently monitor and analyze the documents against (a) the public regulatory framework (including legislation and case law from the EU and Member States, public provisions and suggestions by authorities, etc.) and (b) private contracts. The system will actively inform companies, persons in charge (directors, managers, data protection officers, etc.) or the company's lawyer(s) whenever there are updates in relevant legislation, case law, or in contractual obligations that affect a company's obligations, even across different jurisdictions and languages.

The *Labour Law* use case provides access to aggregated and interlinked legal information regarding labour law across

multiple legal orders, jurisdictions, and languages. The prototype analyses labour legislation from the EU and Member States, and jurisprudence related to labour law issues. This use case makes use of lists of Frequently Asked Questions (FAQ) regarding employment and labour relations, which should be privileged whenever looking for answers posed in natural language. The platform addresses the integration of these heterogeneous documents, coming from different jurisdictions, in various languages, with unequal structure, temporal validity, and geographical scope, which will ultimately benefit the Digital Single Market.

The *Oil and Gas* use case is focused on compliance management support for geothermal energy projects and aims to obtain standards and regulations associated with certain terms in the field of geothermal energy. A user can submit a Request For Proposal (RFP) or feasibility study to the system and is then informed about which standards or regulations must be taken into consideration to carry out the considered project in a compliant manner. This scenario will innovate and speed up existing compliance related services. The uploaded user documents (RFP or feasibility study) are analyzed and, with the help of semantic services (Section 4), the most important terms are identified. The terms of interest include geolocations, types of activities, types of machinery involved, names of organizations, possible mentions of relevant regulations. Next up, the prototype uses the search service to find the documents in the LKG that are most relevant to the uploaded user document. The ranking of the retrieved documents is executed by the semantic similarity service. Thanks to machine translation, the prototype is able to deal with multiple languages – the documents presented to the user are not necessarily in the same language as the uploaded document.

3. Legal Knowledge Graph

Knowledge graphs represent entities as nodes, attributes as node labels and the relationship between entities as edges. RDF is particularly well suited for representing knowledge graphs, and, indeed, the recent attention has finally brought Semantic Web technologies back into the centre of current research and development trends after years of silent existence. Many AI and NLP applications rely on knowledge graphs as crucial resources, such as, among others, information search, data integration, data analytics, question answering and context-dependent recommendations.

In the multilingual legal domain, knowledge graphs have the full support of public institutions, which are publishing massive amounts of linked data, which are becoming critical assets of the companies operating them. The amount of legal data made accessible either in free or for-a-fee modalities by legal information providers can be hardly imagined. In 2014, Lexis Nexis claimed to have 30 Terabytes of content, WestLaw accounted for more than 40,000 databases[2]. Their value can be roughly estimated: as early as 2012, the four big players (WestLaw, Lexis Nexis, Wolters Kluwer and Bloomberg Legal) about US-$10,000M in total revenue. Language data (e. g., resources with any kind of linguistic information) belongs to a much smaller domain, but is still, unmanageable as a whole.

We are interested in a small fraction of the information belonging to these domains. In particular, Lynx is using the data necessary to provide the compliance services (Section 2) – the earliest Lynx knowledge graph being the Spanish legislation enriched with annotations (Rodríguez-Doncel et al., 2018). As shown in Figure 1, the scope of the data in the Lynx Multilingual Legal Knowledge Graph is legal and regulatory data (mainly comprising legislation, case law and standards-related data), on the one hand, and language data (such as corpora, terminologies, glossaries or dictionary data), on the other, to cover the multilingual aspects of the services. The Lynx platform will try to comprehensively identify every possible open dataset in the intersection of these domains as its core category.

Figure 1 shows the core of the Lynx Multilingual Legal Knowledge Graph in the strict sense, as the set of entities whose URI is within the 'lynx' top level domain, but with links to external entities in the wider Web of Data. The definitions of both language data and regulatory data are fuzzy, but flexible as to introduce data of many different kinds whenever necessary (geographical data, user information, etc.). Because data in the Semantic Web can not be separated from the data models, and data models are accessed in the same manner as data, ontologies and vocabularies are part of the LKG as well. Moreover, any kind of metadata (describing documents, standards etc.) is also part of the LKG, as well as the description of the entities producing the documents (courts, users, jurisdictions). In order to provide the compliance services, and with different degree of interest, both primary and secondary law are of use, and any relevant document in a wide sense may become part of the Legal Knowledge Graph.

4. Lynx Platform

The base infrastructure of the Lynx platform follows the paradigm of a microservice architecture, which is a variant of the service-oriented architecture (SOA) where an application is structured as a collection of loosely coupled services. Each microservice can be developed and deployed independently, which also allows the use of different programming languages for their implementation. Microservices are small and autonomous and can be developed more efficiently than monolithic, integrated systems. In addition, the deployment of microservices can, to a very large extent, be automated, also facilitating the monitoring of individual services. A crucial advantage is concerned with the scalability of systems based on microservices, which is a lot easier than scaling monolithic systems. The communication between services is based on REST interfaces, allowing the simple communication and decoupling of the client from the server. We decided to make use of containers (specifically Docker[3]) and an architecture that can host and manage several containers (OpenShift[4], a containerization software built on top of Kubernetes[5]).

[2]LexisNexis Legal and Professional, see http://lexisnexis.com

[3]https://www.docker.com

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

[5]https://kubernetes.io

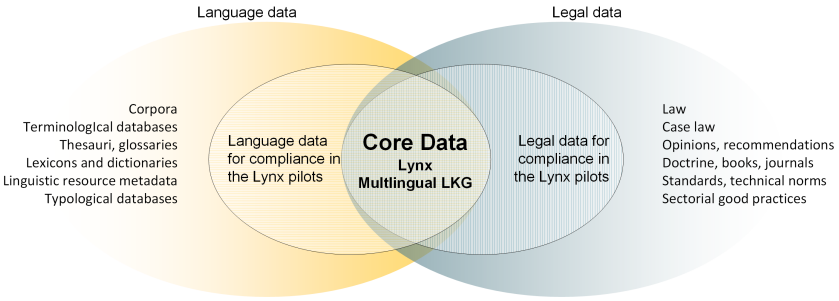

Figure 1: Scope of the Core Data in the Multilingual Legal Knowledge Graph

4.1. Semantic Services

In the following we describe the semantic processing services. This is a heterogeneous set: some of the services make use of others, some extract or annotate information (e. g., NER or Temporal Expression Analysis), while others operate on full documents, yet others provide a user interface (e. g., QA). A complete description is out of scope for this paper and can be found in Rehm et al. (2019).

Term Extraction Enables the creation of a taxonomy for a certain use case, domain or company, using the cloud-based Tilde Terminology service[6]. It extracts terms following the methodology by (Pinnis et al., 2012), creating a SKOS vocabulary containing terms, contexts and references to their source documents.

Linguistic Resources The LKG has to be adaptable across domains and sectors. It is based on a collection of domain-dependent and domain-independent vocabularies accessible through a common RDF graph. The domain-dependent vocabularies comprise terminologies coming from the legal sector and the use case domains (e. g., EuroTermBank[7]), and others created from scratch to cover the specific needs of the business cases, taking advantage of the Linguistic Linked Open Data Cloud (LLOD). The domain-independent vocabularies are taken from lexicographic data published by KDictionaries.[8] They contain cross-lingual links for the five languages served by our platform (Dutch, English, German, Italian, Spanish). Besides their overall coverage of solely domain-independent vocabularies, they contain information on words and phrases that include also or only domain-dependent meanings (e. g., court for the former, lawyer for the latter). Domain-independent dictionary data provide a common ground across domains that facilitates traversing semantically annotated documents coming from different specialised domains (e. g., Legal or Oil & Gas). They also support certain NLP functionalities such as Word Sense Disambiguation by providing a common catalogue of word senses. The data is being remodeled in RDF according to the Ontolex Lemon Lexicography Module Specification.[9] Right now that linguistic information is being used by other services, WSD, Search, QADoc, to get synonyms, term variants and translations that help in the cross-lingual search and cross-lingual question answering.

Named Entity Recognition The NER service (Leitner et al., 2019; Leitner et al., 2020) is based on several trained models, Conditional Random Fields (CRFs) and bidirectional Long-Short Term Memory Networks (BiLSTMs) (Huang et al., 2015; Lample et al., 2016; Riedl and Padó, 2018). This service includes an entity linking module in which we retrieve a unique identifier (URI) for the spotted entities. It uses DBPedia SPARQL[10] and DBPedia spotlight.[11] Retrieved URIs are stored as part of the entity annotation.

Concept Extraction Concept extraction enables the insertion of links between documents and elements of controlled vocabularies in the LKG. These relations are the first step for enriching text fragments with knowledge from the LKG. Importantly, the inclusion of labels in many languages allows linking of documents in different languages, combining the knowledge derived from them, as well as multilingual search and recommendation. The Concept Extraction service works in as many languages as the taxonomies have labels in, and thus we can leverage multinational efforts for creating multilingual taxonomies such as EUROVOC[12] or UNBIS[13].

Word Sense Disambiguation To enable the use of incomplete KGs for automatic text annotations, we introduce a robust method for discriminating word senses using thesaurus information like hypernyms, synonyms, types/-classes, contained in the KG (Revenko and Mireles, 2017). It uses collocations to induce word senses and to discriminate the thesaurus sense from others. The service is used for any kind of entity linking, especially after NER, to correctly identify which named entities are indeed within the vocabulary scope of the LKG.

Temporal Expression Analysis A prototype analyses time expressions in German-language legal documents, especially court decisions and legislative texts. The annotation of temporal expressions is important, but the most interesting part is the normalization, which can be used for interlinking documents (or parts of documents) using the normalized values of temporal expressions.

Legal Reference Resolution Usually, editors attempt to be consistent and follow patterns to reference other documents. The developed methodology, currently imple-

[6]https://term.tilde.com
[7]http://www.eurotermbank.com
[8]https://www.lexicala.com
[9]http://www.w3.org/ns/lemon/lexicog

[10]https://dbpedia.org/sparql
[11]https://www.dbpedia-spotlight.org
[12]https://publications.europa.eu/en/web/eu-vocabularies/
[13]http://metadata.un.org/?lang=en

mented as a language-agnostic prototype, follows this assumption and attempts to discover patterns used in a semi-automatic manner. Patterns are constructed from features that are either individual tokens (e. g., "Decision", "EU", etc.) or processed features (e. g., "DIGITS" as a placeholder for numbers).

Semantic Similarity We use a hybrid type of similarity measure. First, the text of the document is annotated, such as the resolution of temporal or geographical references. Second, similarity is computed using a linear combination of text-based and knowledge-based similarities. The former are encoded by cosine-similarity of TF-IDF vectors and the latter by the overlap as measured by Jaccard coefficient of entities that the two documents either mention directly, or are linked in the LKG to mentioned ones. This approach allows us to detect similarity between documents even if they have only few entities in common.

Question Answering The Question Answering (QA) service accepts a natural language question and responds with an answer, extracted from a document in a given corpus. The end-to-end system consists of three components: 1) The Query Formulation module transforms a question into a query, which can be expanded using a domain specific vocabulary from the LKG. The query is processed through an indexer to obtain matching documents from the corpora. 2) The Answer Generation module extracts potential answers from the retrieved documents from the LKG. 3) The Answer Selection module identifies the best answer based on various criteria such as local structure of the text and global interaction between each pair of words based on specific layers of the model.

4.2. Document Manager

The Document Manager (DCM) forms a central part of the Lynx platform in terms of the general platform capabilities; this is where documents are stored and maintained. Its basic functionality includes the storage and annotation of documents with an emphasis on keeping them synchronized, providing read and write access, as well as updates of documents and annotations.

The DCM can be queried in terms of annotations (e. g., "which documents contain mentions of this entity?"), and in terms of documents (e. g., "what are the contents/annotations of document X?"). All queries to the DCM are executed through REST. The interface includes a set of Create, Read, Update, and Delete (CRUD) APIs to manage collections, documents and annotations within the Lynx platform. Through their representation in JSON-LD (namely, RDF), Lynx documents are not only isolated elements but nodes in a graph. The use of semantics to formalize the meaning of the classes and properties qualifies this graph to be called an actual Knowledge Graph. The DCM is implemented as a Linked Data Platform (LDP) server based on Trellis[14], which is why basic metadata about a document is stored as triples natively – our implementation is based on Elastic Search and stores a JSON-LD serialisation of RDF. Document structure information and various types of metadata such as subject, jurisdiction, language etc. are also triplified through the DCM at storing time. The NIF (Hellmann et al., 2012) Version 2.1 ontology is used to describe the structure metadata and a mashup of metadata-specific ontologies are used for other descriptive, structural or administrative metadata. Annotations of each document are also described using NIF V2.1. An overview of the Lynx data model can be found online.[15] Triples from all documents including data and metadata can be queried using the SPARQL endpoint provided by Trellis, thus providing access to the LKG including the ability to evaluate complex queries – the equivalent for the ElasticSearch implementation being made by periodic data exports, queryable through the endpoint.[16] Extensive usage of vocabularies as values for metadata or annotations increases the value of the LKG and the interoperability of the system. The DCM is the main building block of the Lynx Legal Knowledge Graph (LKG), it is where the LKG resides. Its basic architecture and core functionalities are described in (Maganza and Anagbo, 2019; Moreno-Schneider and Rehm, 2018c).

4.3. Workflow Manager

Using a microservice architecture enforces the use of some kind of management tool in order to orchestrate the execution of the different services involved in more complex tasks (Maganza and Anagbo, 2019). The combination of several functionalities from different services is defined as a workflow and the module responsible for orchestrating them is called workflow manager (WM). Our previous work includes a generic workflow manager for curation technologies (Bourgonje et al., 2016a), and two indicative descriptions of the initial prototype of the Lynx workflow manager (Moreno-Schneider and Rehm, 2018a; Moreno-Schneider and Rehm, 2018b). The final Lynx workflow manager is based on the Camunda BPMN engine[17] because Camunda was in a more mature state than any of the alternatives. The requirements of the WM are presented in (Moreno-Schneider and Rehm, 2018c), while the Lynx workflows are described in (Moreno-Schneider and Rehm, 2018d; Moreno-Schneider and Rehm, 2019). Figure 2 shows the architecture of the workflow manager. Its main components are described in the following sections.

4.3.1. Workflow Manager Engine

The Workflow Manager Engine (WME) is responsible for converting workflows into tasks for the workers. It is based on Camunda. The main concepts of this component are:

- Workflow: a direct acyclic graph whose nodes are associated with tasks;
- Task: an atomic unit of business logic, a task is associated with one and only one Lynx peripheral service;
- Process: a runtime instance of a workflow;
- Job: a runtime instance of a task.

The WME also provides a complete REST interface for managing workflow executions and templates. Apart from

[14]https://github.com/trellis-ldp/trellis

[15]http://lynx-project.eu/data2/data-models

[16]http://sparql.lynx-project.eu

[17]https://camunda.com

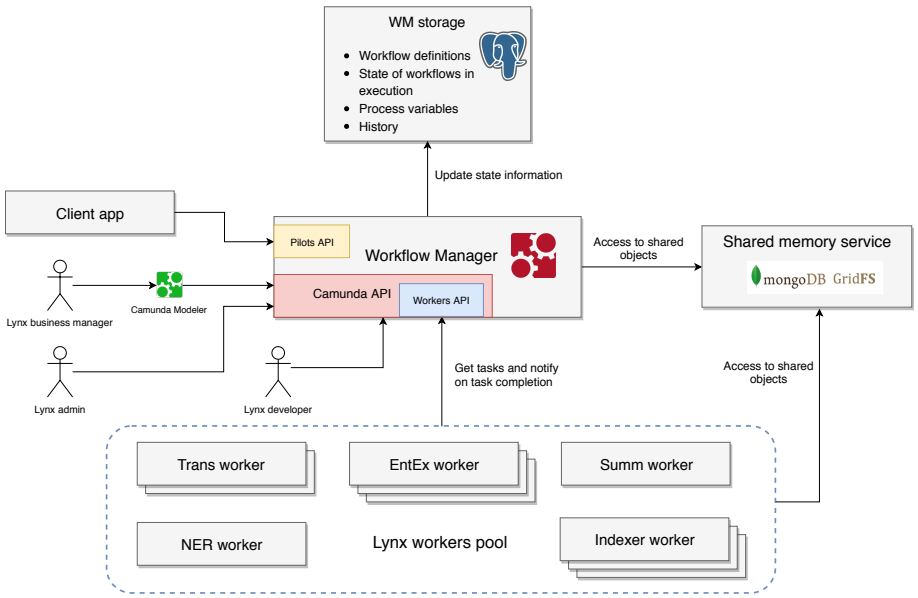

Figure 2: Workflow Manager Architecture

that, the WME uses some internal storage to store the different objects (workflows, tasks, processes and jobs) that are created during execution. All these elements are stored in the Workflow Manager Storage, which is implemented using a PostgreSQL[18] database.

4.3.2. Workers

The Workers are responsible for the execution of tasks inside a workflow. Every task is identified by a topic name like "TimEx-LKGPopulation" or "NER-ContractAnalysis". Every worker uses the Camunda External Task Client library[19] to connect to the WME to obtain the tasks it has to execute. Currently, there are four implemented types of workers. Each type can be instantiated multiple times with different configuration files, i. e., each instantiated worker is responsible (based on the configuration) to connect to a different service, or even to the same service but with different parameters (e. g., "lang=de" or "lang=en"). The four types of workers have different functionalities:

(i) the *document-translation-worker* connects to the Tilde translation services;

(ii) the *document-enrichment-worker* connects to one of the enrichment services inside the Lynx platform (NER, TIMEX, SUMM, WSID, EntEx, etc.);

(iii) the *save-enriched-doc-in-LKG-worker* saves an enriched document inside the DCM; and

(iv) the *create-enriched-document-worker* creates the enriched document. It collects all annotations produced by the services and aggregates them to create a Lynx document. This document is stored in the DCM or sent back to the client.

4.3.3. Shared Memory Service

The shared memory service is used by both the workflow manager engine and the workers to share large data objects. For instance, the WM uses it to share with the workers the documents they have to process. The shared memory service uses a MongoDB[20] database.

4.3.4. Pilot API

The Pilot API is a component of the WM and responsible for accessing, managing and executing the workflows of the project pilots. It consists of four discrete methods: three HTTP POST methods to execute and manage specific workflows (one method per pilot); and one HTTP GET method to retrieve the current state of a concrete workflow.

4.3.5. Graphical User Interface

Workflows are described using the BPMN (Business Process Model and Notation) standard (OMG, 2011). Considering that specifying BPMN files manually is not the most user friendly approach, we integrated a graphical user interface for the definition of new workflows. We decided to use the Camunda Modeler.[21]

4.4. Defined Workflows

We conceptualise the requirements of the different use cases as content curation workflows (Moreno-Schneider and Rehm, 2018b; Bourgonje et al., 2016a; Bourgonje et al., 2016b; Rehm et al., 2018). Workflows are defined as the execution of specific services to perform the processing of one or more documents under the umbrella of a certain task or use case. The specification of a workflow includes its input and output as well as the functionality it is supposed to perform: annotate or enrich a document, add a document to the knowledge base, search for information, etc. A workflow makes use of one or more service to implement a required functionality. For the definition of the workflows we

[18] https://www.postgresql.org
[19] https://docs.camunda.org/manual/7.9/user-guide/ext-client/
[20] https://www.mongodb.com
[21] https://camunda.com/download/modeler/

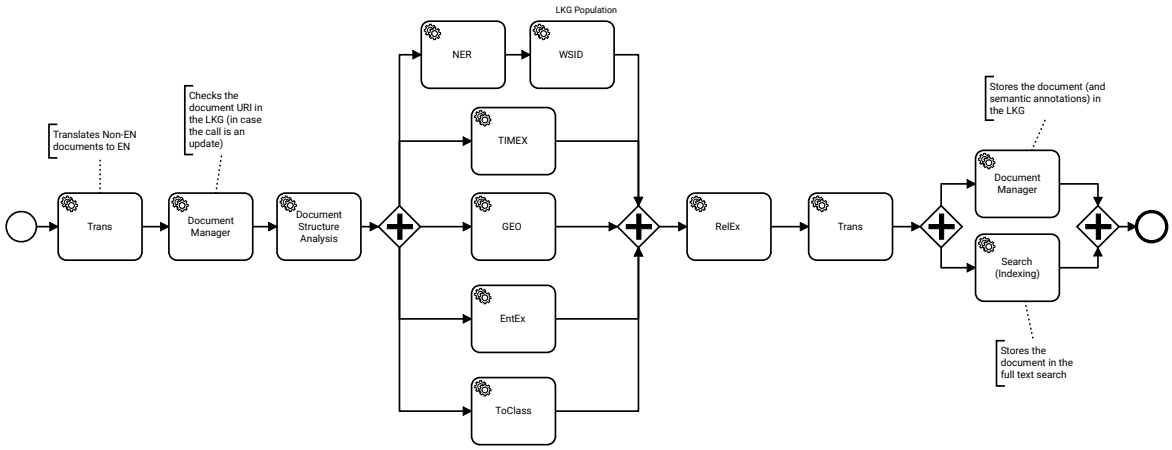

Figure 3: Legal Knowledge Graph population workflow

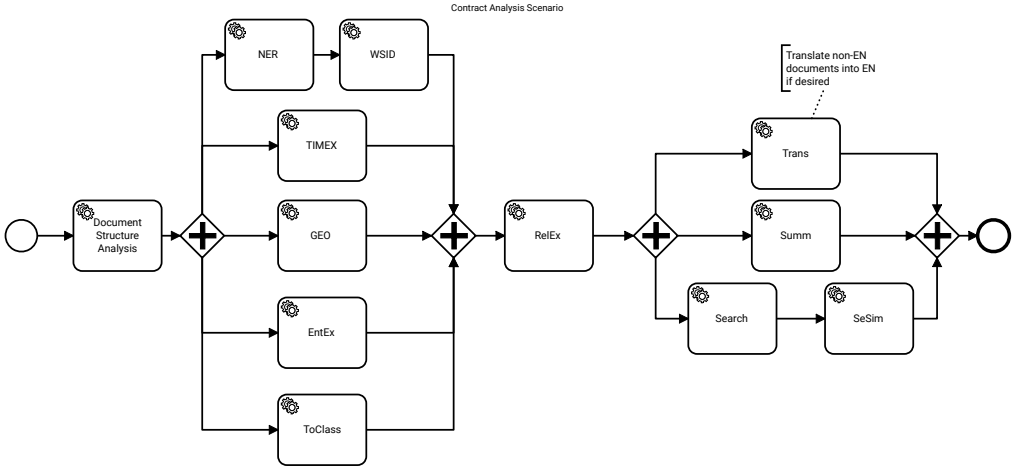

Figure 4: Contract analysis workflow

performed a systematic analysis of the services, developed in parallel, and matched them with the required functionalities. First, we determine the principal elements involved, i. e., the services, input and output. Second, we define the order in which the services have to be executed. Third, we identify the components shared between workflows.

Currently we have defined three different workflows: (1) *Legal Knowledge Graph Population* is responsible for the initial population of the LKG (Figure 3) by semantically annotating and then storing documents in the LKG; (2) *Contract Analysis* for the processing, analysis and enrichment of contracts (Figure 4); (3) *Geothermal Project Analysis* is responsible for the analysis of geothermal project proposals in order to check their compliance with the applicable regulations.

5. Related Work

There are several systems, platforms and approaches that are related to the technology platform, which is under development in our project. In the wider area of legal document processing, technologies from several fields are relevant, among others, knowledge technologies, citation analysis, argument mining, reasoning and information retrieval. Literature overviews can be found in (Moreno-Schneider and Rehm, 2018a) and (Agnoloni and Venturi, 2018).

Commercial Systems and Services – The LexisNexis system is the market leader in the legal domain; it offers services, such as legal research, practical guidance, company research and media-monitoring as well as compliance and due diligence. WestLaw is an online service that allows legal professionals to find and consult relevant legal information.[22] One of its goals is to enable professionals to put together a strong argument. There are also smaller companies that offer legal research solutions and analytic environments, such as RavelLax,[23] or Lereto[24]. A commercial search engine for legal documents, iSearch, is a service offered by LegitQuest.[25] The Casetext CARA Research Suite allows uploading a brief and then retrieving, based on its contents, useful case law.[26] There is also a growing number of startup companies active in the legal domain. All these systems are commercial, therefore you have to pay for their use. Our platform, on the other hand, does not have a com-

[22] http://legalsolutions.thomsonreuters.com/law-products/westlaw-legal-research/

[23] http://ravellaw.com

[24] https://www.lereto.at

[25] https://www.legitquest.com

[26] https://casetext.com

mercial base and you do not have to pay for it, only some of the services that are available under specific licenses, for which you would have to pay.

Research Prototypes – Most of the documented research prototypes were developed in the 1990s under the umbrella of Computer Assisted Legal Research (CALR) (Span, 1994). In the following we briefly review several of these systems, which usually focus on one very specific feature or functionality. One example is the open source software for the analysis and visualisation of networks of Dutch case law (van Kuppevelt and van Dijck, 2017). This technology determines relevant precedents (analysing the citation network of case law), compares them with those identified in the literature, and determines clusters of related cases. A similar prototype is described by (Agnoloni et al., 2017). (Gifford, 2017) propose a search engine for legal documents where arguments are extracted from appellate cases and are accessible through selecting nodes in a litigation issue ontology or relational keyword search. Lucem (Bhullar et al., 2016) mirrors the way lawyers approach legal research, developing visualisations that provide lawyers with an additional tool to approach their research results. The Eunomos prototype semi-automates the construction and analysis of knowledge (Boella et al., 2012). The main difference between all these tools and our platform is the type of documents they work with. Most of these systems are limited to a single type of document, while we work with a wide variety, from contracts or laws (labour law) to industrial standards. In addition, each of these tools has a specific functionality, while the Lynx platform combines them all in a single ecosystem.

6. Summary and Future Work

We present the technology platform currently under development in the project Lynx, focusing upon curation workflows and processing services. These serve two main purposes: 1) to extract semantic information from large and heterogeneous sets of documents to ingest the extracted information into the Legal Knowledge Graph; 2) to extract semantic information from documents that users of the platform work with. In addition to the semantic extraction, we provide services for the processing and curation of whole documents with the goal of mapping extracted terms and concepts to the LKG, and services that aim at accessing the LKG (question answering). The final prototypes and the whole platform will be available during the last months of the project, starting from summer 2020.

Future work includes the completion of service development, adapting the services to all languages required in the project's use cases, implementing the pilot applications and developing the web interface of the platform. In addition, we will finalise, deploy and evaluate the workflow manager and workflows defined in the project. This will not only improve the performance of the system but also simplify the way users can access the system. Last but not least, as an additional exploitation option we will explore the integration and deployment of the Lynx services through the European Language Grid (Rehm et al., 2020).

While some of the technologies and data sets developed in Lynx are proprietary, the following will be made openly available at the end of the project at the very latest: Legal NER for German (models and data sets) (Leitner et al., 2020), Temporal Expression Analyzer for Spanish and English[27] (Navas-Loro, 2017; Navas-Loro and Rodríguez-Doncel, 2019), Word Sense Induction and Disambiguation for English[28] (Revenko and Mireles, 2017), the WME and the DCM among others.

Acknowledgments

This work has been partially funded by the project LYNX, which has received funding from the EU's Horizon 2020 research and innovation programme under grant agreement no. 780602, see http://www.lynx-project.eu.

7. References

Agnoloni, T. and Venturi, G. (2018). Semantic processing of legal texts. In Jacqueline Visconti, editor, *Handbook of Communication in the Legal Sphere*, pages 109–138. De Gruyter, Berlin, Boston.

Agnoloni, T., Bacci, L., Peruginelli, G., van Opijnen, M., van den Oever, J., Palmirani, M., Cervone, L., Bujor, O., Lecuona, A. A., García, A. B., Caro, L. D., and Siragusa, G. (2017). Linking european case law: BO-ECLI parser, an open framework for the automatic extraction of legal links. In Wyner and Casini (Wyner and Casini, 2017), pages 113–118.

Bhullar, J., Lam, N., Pham, K., Prabhakaran, A., and Santillano, A. J., (2016). *Lucem: A Legal Research Tool*. Number 63. Computer Engineering Senior Theses.

Boella, G., di Caro, L., Humphreys, L., Robaldo, L., and van der Torre, L. (2012). Nlp challenges for eunomos, a tool to build and manage legal knowledge.

Bourgonje, P., Moreno-Schneider, J., Nehring, J., Rehm, G., Sasaki, F., and Srivastava, A. (2016a). Towards a Platform for Curation Technologies: Enriching Text Collections with a Semantic-Web Layer. In Harald Sack, et al., editors, *The Semantic Web*, number 9989 in Lecture Notes in Computer Science, pages 65–68. Springer, June. ESWC 2016 Satellite Events. Heraklion, Crete, Greece, May 29 – June 2, 2016 Revised Selected Papers.

Bourgonje, P., Schneider, J. M., Rehm, G., and Sasaki, F. (2016b). Processing Document Collections to Automatically Extract Linked Data: Semantic Storytelling Technologies for Smart Curation Workflows. In Aldo Gangemi et al., editors, *Proceedings of the 2nd International Workshop on Natural Language Generation and the Semantic Web (WebNLG 2016)*, pages 13–16, Edinburgh, UK, September. The Association for Computational Linguistics.

Chalkidis, I. and Androutsopoulos, I. (2017). A deep learning approach to contract element extraction. In *JURIX*.

Gifford, M. (2017). Lexridelaw: an argument based legal search engine. In *ICAIL '17*.

Hellmann, S., Lehmann, J., and Auer, S. (2012). Nif: An ontology-based and linked-data-aware nlp interchange format. *Working Draft*, page 252.

[27] https://github.com/mnavasloro/Annotador

[28] https://github.com/semantic-web-company/ptlm_wsid

Huang, Z., Xu, W., and Yu, K. (2015). Bidirectional LSTM-CRF models for sequence tagging. *CoRR*, abs/1508.01991.

Lample, G., Ballesteros, M., Subramanian, S., Kawakami, K., and Dyer, C. (2016). Neural architectures for named entity recognition. In *NAACL HLT 2016, The 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, San Diego California, USA, June 12-17, 2016*, pages 260–270.

Leitner, E., Rehm, G., and Moreno-Schneider, J. (2019). Fine-grained Named Entity Recognition in Legal Documents. In Maribel Acosta, et al., editors, *Semantic Systems. The Power of AI and Knowledge Graphs. Proceedings of the 15th International Conference (SEMANTiCS 2019)*, number 11702 in Lecture Notes in Computer Science, pages 272–287, Karlsruhe, Germany, 9. Springer. 10/11 September 2019.

Leitner, E., Rehm, G., and Moreno-Schneider, J. (2020). A Dataset of German Legal Documents for Named Entity Recognition. In Nicoletta Calzolari, et al., editors, *Proceedings of the 12th Language Resources and Evaluation Conference (LREC 2020)*, Marseille, France, May. European Language Resources Association (ELRA). Accepted for publication.

Maganza, F. and Anagbo, K. J. (2019). Lynx D1.4 Setup and implementation of the basic platform, May.

Moreno-Schneider, J. and Rehm, G. (2018a). Curation Technologies for the Construction and Utilisation of Legal Knowledge Graphs. In Georg Rehm, et al., editors, *Proceedings of the LREC 2018 Workshop on Language Resources and Technologies for the Legal Knowledge Graph*, pages 23–29, Miyazaki, Japan, May. 12 May 2018.

Moreno-Schneider, J. and Rehm, G. (2018b). Towards a Workflow Manager for Curation Technologies in the Legal Domain. In Georg Rehm, et al., editors, *Proceedings of the LREC 2018 Workshop on Language Resources and Technologies for the Legal Knowledge Graph*, pages 30–35, Miyazaki, Japan, May. 12 May 2018.

Moreno-Schneider, J. and Rehm, G. (2018c). Lynx D4.1 Pilots Requirements Analysis Report, May.

Moreno-Schneider, J. and Rehm, G. (2018d). Lynx D4.2 Initial version of Workflow Definition, November.

Moreno-Schneider, J. and Rehm, G. (2019). Lynx D4.3 Final version of Workflow Definition, May.

Navas-Loro, M. and Rodríguez-Doncel, V. (2019). Annotador: a Temporal Tagger for Spanish.

Navas-Loro, M. (2017). Mining, Representation and Reasoning with Temporal Expressions in the Legal Domain. In *Doctoral Consortium, Challenge, Industry Track, Tutorials and Posters @ RuleML+RR*.

OMG. (2011). Business Process Model and Notation (BPMN), Version 2.0, January.

Pinnis, M., Ljubešić, N., Ştefănescu, D., Skadiņa, I., Tadić, M., and Gornostay, T. (2012). Term Extraction, Tagging, and Mapping Tools for Under-Resourced Languages. In *Proceedings of the 10th Conference on Terminology and Knowledge Engineering (TKE 2012)*, number June 2012, pages 193–208, Madrid, Spain.

Rehm, G., Schneider, J. M., Bourgonje, P., Srivastava, A., Fricke, R., Thomsen, J., He, J., Quantz, J., Berger, A., König, L., Räuchle, S., Gerth, J., and Wabnitz, D. (2018). Different Types of Automated and Semi-Automated Semantic Storytelling: Curation Technologies for Different Sectors. In Georg Rehm et al., editors, *Language Technologies for the Challenges of the Digital Age: 27th International Conference, GSCL 2017, Berlin, Germany, September 13-14, 2017, Proceedings*, number 10713 in Lecture Notes in Artificial Intelligence (LNAI), pages 232–247, Cham, Switzerland, January. Gesellschaft für Sprachtechnologie und Computerlinguistik e.V., Springer. 13/14 September 2017.

Rehm, G., Moreno-Schneider, J., Gracia, J., Revenko, A., Mireles, V., Khvalchik, M., Kernerman, I., Lagzdins, A., Pinnis, M., Vasilevskis, A., Leitner, E., Milde, J., and Weißenhorn, P. (2019). Developing and Orchestrating a Portfolio of Natural Legal Language Processing and Document Curation Services. In *Proceedings of Workshop on Natural Legal Language Processing (NLLP 2019)*, Minneapolis, USA, June. Co-located with NAACL 2019. 7 June 2019.

Rehm, G., Berger, M., Elsholz, E., Hegele, S., Kintzel, F., Marheinecke, K., Piperidis, S., Deligiannis, M., Galanis, D., Gkirtzou, K., Labropoulou, P., Bontcheva, K., Jones, D., Roberts, I., Hajic, J., Hamrlova, J., Kacena, L., Choukri, K., Arranz, V., Vasiljevs, A., Anvari, O., Lagzdins, A., Melnika, J., Backfried, G., Dikici, E., Janosik, M., Prinz, K., Prinz, C., Stampler, S., Thomas-Aniola, D., Perez, J. M. G., Silva, A. G., Berrio, C., Germann, U., Renals, S., and Klejch, O. (2020). European Language Grid: An Overview. In Nicoletta Calzolari, et al., editors, *Proceedings of the 12th Language Resources and Evaluation Conference (LREC 2020)*, Marseille, France, May. European Language Resources Association (ELRA). Accepted for publication.

Revenko, A. and Mireles, V. (2017). Discrimination of word senses with hypernyms. In Anna Lisa Gentile, et al., editors, *Proceedings of the 5th International Workshop on Linked Data for Information Extraction co-located with the 16th International Semantic Web Conference (ISWC 2017), Vienna, Austria, October 22, 2017.*, volume 1946 of *CEUR Workshop Proceedings*, pages 50–61. CEUR-WS.org.

Riedl, M. and Padó, S. (2018). A named entity recognition shootout for german. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, volume 2, pages 120–125. Association for Computational Linguistics.

Rodríguez-Doncel, V., Navas-Loro, M., Montiel-Ponsoda, E., and Casanovas, P. (2018). Spanish legislation as linked data. In *Proceedings of the 2nd Workshop on Technologies for Regulatory Compliance co-located with the 31st International Conference on Legal Knowledge and Information Systems (JURIX 2018), Groningen, The Netherlands, December 12, 2018.*, pages 135–141.

Span, G. (1994). Lites: An intelligent tutoring system shell for legal education. *International Review of Law, Com-*

puters & Technology, 8(1):103–113.

van Kuppevelt, D. and van Dijck, G. (2017). Answering legal research questions about dutch case law with network analysis and visualization. In Wyner and Casini (Wyner and Casini, 2017), pages 95–100.

Adam Z. Wyner et al., editors. (2017). *Legal Knowledge and Information Systems – JURIX 2017: The Thirtieth Annual Conference, Luxembourg, 13-15 December 2017*, volume 302 of *Frontiers in Artificial Intelligence and Applications*. IOS Press.