



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

D5.5 Demonstrator for pilot 1

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	Christian Sageder (Cybly, former openlaws)
CONTRIBUTORS	
REVIEWERS	Pascual Boil Ballesteros (CC), Pieter Verhoeven (DNV.GL), Socorro Bernardos (UPM), Víctor Rodríguez-Doncel (UPM), Elena Montiel-Ponsoda (UPM), María Navas-Loro (UPM)
VERSION STATUS	V 1.0 Final
NATURE	Demonstrator
DISSEMINATION LEVEL	Public
DOCUMENT DOI	10.5281/zenodo.4298949
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	11/11/2020	Christian Sageder (Cybly / former openlaws)
02	Review	17/11/2020	1 st review by Pieter Verhoeven (DNV.GL), María Navas-Loro (UPM) and Pascual Boil Ballesteros (CC)
03	2 nd draft, additional content, Proofreading	26/11/2020	Christian Sageder (Cybly / former openlaws), Ryan, Michael (Cybly)
04	3 rd draft, updated screenshots and drawings	29/11/2020	Christian Sageder
05	Some additions regarding missing information on how to get access to the pilot	30/11/2020	Christian Sageder
1.0	Final review	30/11/2020	Christian Sageder

DISCLAIMER

This document does not represent the opinion of the European Community, and the European Community is not responsible for any use that might be made of its content. Neither the Lynx consortium as a whole, nor a certain party of the Lynx consortium warrants that the information contained in this document is capable of use, nor that using the information is free from risk, and does not accept any liability for loss or damage suffered by any person using this information.

EXECUTIVE SUMMARY

The main purpose of this document is to give an insight into the development of pilot 1 (contract management) which is being developed within the Lynx project.

Contracting is a common activity in companies, but managing contracts systematically is a cumbersome activity only few companies are effective at. Contracts and relevant documents are physically and electronically distributed across the entire organization. Against this backdrop, we are focusing on automated contract analysis to provide intelligent contract archiving solutions and compliance services.

In order to achieve this, the following solutions have been implemented:

- A back-end solution to analyze single and multiple contracts including an archive which can be used by other services / applications.
- A front-end solution where a single user has the possibility to manage contracts.
- A demo website where a single document can be analyzed

For the front-end solution selected provisional screenshots and descriptions are provided. In addition, a description of the architecture and newly developed services for the back-end is given.

This deliverable is structured as follows: The first section deals with the relevance of contract management and embeds the efforts into the broader Lynx context. Section 2 describes several use cases we aim at solving with this pilot. In light of this, we are discussing objectives, our working basis in the pilot phase. Section 3 describes the pilot development in relation to the objects and Section 4 (pilot architecture) contains information about the underlying back-end solutions. Finally, the deliverable closes with a reflection and outlook.

TABLE OF CONTENTS

1	INTRODUCTION	6
2	USE CASES AND OBJECTIVES	7
2.1	Objectives	7
2.2	Target Group	8
2.3	General Description	8
2.4	Visualization of basic modules	9
2.5	Current stage of the pilot	13
2.6	working basis in the pilot	14
3	Pilot Description	15
3.1	Harvesting	15
3.2	Conversion	15
3.3	Extraction	17
3.4	Archiving	17
3.5	Search	17
3.6	Manage	19
3.7	Update	19
3.8	Export	20
4	ARCHITECTURE	21
4.1	Technology	¡Error! Marcador no definido.
4.2	DirWatch	21
4.3	BROKER SerVICE	22
4.4	Document Converter Service	23
4.5	Annotation Service	24
4.6	RegExFinder Service	25
4.7	Lynx Services	25
5	REFLECTIONS AND RECOMMENDATIONS FOR FURTHER DEVELOPMENT	27
5.1	Test Data	27
5.2	Annotations	27
5.3	NLP Models	28
6	References	29

TABLE OF FIGURES

Figure 1 Solution approaches	7
Figure 2. Main page for a single contract	9
Figure 3. Table of Content	10
Figure 4. Detailed View of an Extracted Information Unit (Seller).....	10
Figure 5. Information Query	11
Figure 6. Manually Add Information (Part 1).....	11
Figure 7. Manually Add Information (Part 2).....	12
Figure 8. Manually Add Information (Part 3).....	12
Figure 9. Further Legal Information.....	13
Figure 10. Further Legal Information (Detailed View).....	13
Figure 11 Sample of an index clause	14
Figure 12 Demo Document-Upload.....	16
Figure 13 Demo Document-View	17
Figure 14 Search with search result.....	18
Figure 15 Search result, with folder view	18
Figure 16 Adding documents to a folder	19
Figure 17 Manage annotation	20
Figure 18. Back-end components	21
Figure 19. DirWatcher.....	22
Figure 20. Sequence when a new document is added.....	23
Figure 21 Annotation Service	24
Figure 22 Sample definition for regex pattern	25

LIST OF TABLES

Table 1. Supported document formats	15
---	----

1 INTRODUCTION

Contracting is a common activity in companies, but managing contracts systematically, which means keeping track of changes or updates, is a cumbersome activity only few companies are effective at. In addition, many SMEs (small and medium enterprise) do not have a database with all the information of their contracts, which prevents them from easily finding information or applying changes.

Let us imagine the following situations in the context of a company:

- a) A contract is needed urgently and no one knows where to find the latest version because the responsible employee left the company. Moreover, the opposing party confronts you with a signed amendment or a subsidiary agreement you've never seen before.
- b) There is a change in law, and you need to know which contracts are affected.
- c) An overview on all obligations with a certain company is needed.

Countless organizations are confronted with similar scenarios, although we are all significantly shaping our legal reality by concluding various contracts. Abstractly, the problem can be summarized as follows: Contracts and contract relevant documents are physically and electronically distributed across the entire organization and tools, e.g. file server, emails, physical documents. As a result, there is often no overview, which leads to inconsistent applications, breaches of contracts and (financial) disadvantages.

The implementation of a comprehensive cross-organizational contract management process appears to be the solution. Flitsch (2010) defines contract management as the creation of ideal structures for:

- contract planning
- contract design
- contract negotiations
- implementation of contracts
- contract administration
- contract archiving

In many cases, organizations are lacking these structures. With this in mind, we are focusing on automated contract administration and archiving. Building on this, we provide smart contract archiving solutions and compliance services. We expect our application to result in enhanced contract compliance, which will ultimately lead to reduced risks and costs for organizations.

All these efforts are embedded into the broader Lynx context and the associated assumption that building a legal knowledge graph - duly interlinked and integrated - results in reducing possible distances to the applicable law and, thus, in facilitating compliant and diligent actions. To this end, we are channeling our efforts to fulfil what Richard Hamming (1962) so aptly formulated decades ago: "The purpose of (scientific) computing is insight, not numbers."

2 USE CASES AND OBJECTIVES

The starting point and, at the same time, the simplest use case is the analysis of a single contract / document. However, reality is much more complex: Regularly, a large number of contracts of diverse nature and purpose needs to be analyzed and kept track of, taking into account various regulatory frameworks. In order to achieve this, we have two approaches. On the one hand, we have a pure back-end solution, and, on the other hand, we are providing a visualization of the created data space for end users. Figure 1 illustrates the different use cases or rather solution approaches in a general manner:

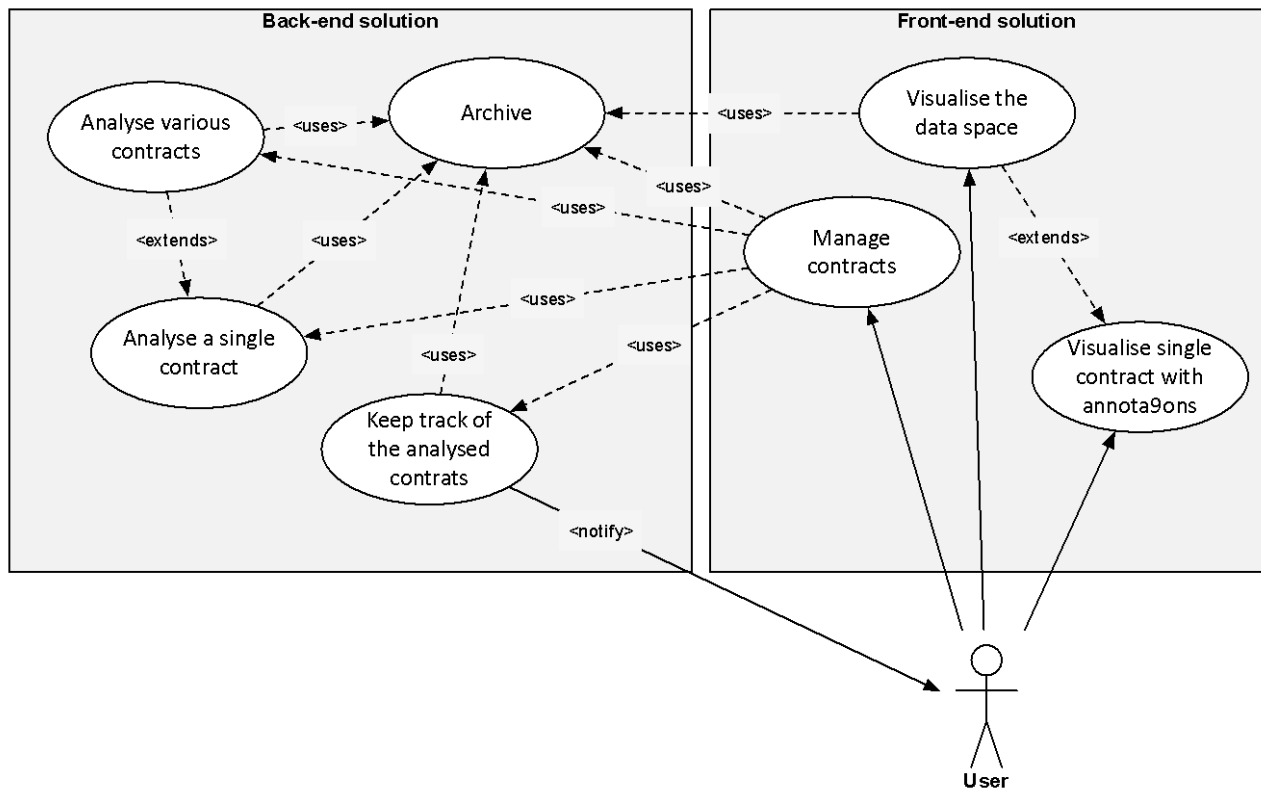


Figure 1 Solution approaches

As core functionality, the back-end solution provides the analysis of single contracts (documents). In addition, multiple contracts can be analyzed. Both functionalities are based on the archive where contracts and all the extracted information is stored. These services can be used by other services / applications.

Through the front-end solution a single user has the possibility to manage (add, delete, update, group, search, etc.) contracts / documents. The user can view a single contract and related annotations or get a broader view of the corresponding data space, e.g. legislation, similar contracts, other contracts with the same partner, etc. In addition, the user is notified when legislation changes with effects on a given contract. All of this supports companies in achieving compliance.

2.1 OBJECTIVES

In the course of the pilot, our goal is to develop reasonable strategies for automated contract analysis and contract archiving. We concentrate on the following key parts of the process (contract administration and management) which are crucial when it comes to defining our application:

- **Harvesting:** There has to be the possibility to harvest documents from different data sources e.g., file system and keeping them up-to-date. Providing an interface for external systems to ingest contract-related documents.
- **Conversion:** Documents in different data formats must be converted to a Lynx-Documents¹, including metadata and document structure. As a minimum, the following formats must be supported: PDF incl. scans, MS-Word, email incl. attachments, Images, e.g. TIFF when they are scans
- **Extraction:** Extracting information from the document by different annotation services.
- **Archiving:** The document and its extracted information must be stored for later reuses. Also, the original document has to be archived, not only the Lynx-Documents representation.
- **Search:** Providing the possibility to search for documents, by full text search, document type, annotations, metadata, e.g. document date
- **Manage:** the possibility to group / restructure the harvested documents
- **Update:** Provide Possibility to add / update annotations and certain meta information
- **Export:** Possibility to export a group of documents

2.2 TARGET GROUP

The intended pilot targets the following groups:

- Law firms
- Medium and Large Enterprises

2.3 GENERAL DESCRIPTION

When it comes to contracts, the tools used are very few. A word processor to create the contracts, email to communicate with the client or the other party and a file storage in a defined directory structure and / or a lawyer software. To keep history, they often add the different version of the document and individual mail communication to this software or put them on the file system.

When it comes to a question like “what is the current contract with X”, they have a look to the file system and pick the most recent version. The same is true for companies, where contracts are also stored somewhere on a server according to a certain structure. For both cases, the assumption is that the latest version has been stored there, what should be the case.

In case there is a change in law and the lawyer wants to inform their customer of this change with an offer to update the contract, e.g. by a side letter, it isn’t always so easy to find all contracts which are affected by this change.

The aim of our solution is not changing the existing workflow, which is well-established in most law firms and companies, but to provide an integrated solution to the existing toolset and workflows

For the user, an advanced search is provided where he/she can search for documents on different facets and where he/she is able to find the documents independently where the documents are stored. In addition, the documents can be organized in multiple structures.

As an additional effect, the targeted solution can also be used for due diligences, as it is possible to organize documents in different structures. These documents can then be exported and made available through external service.

¹ <https://lynx-project.eu/doc/lkg/>

2.4 VISUALIZATION OF BASIC MODULES

In the following section we present selected screenshots of our **visualization of a single contract**, specifically of a purchase contract (“Kaufvertrag zwischen Unternehmern über bewegliche Sachen aus Sicht des Verkäufers”).

Figure 2 shows the main page for a single contract. The top bar contains the contract title as well as relevant metadata, e.g. document title, document date and language. Underneath, the contract is presented on the left-hand side and the headings of extracted information units (in this case about the seller (“Verkäufer”) and the buyer (“Käufer”)) on the right-hand side. Further information units are added based on the contract type.

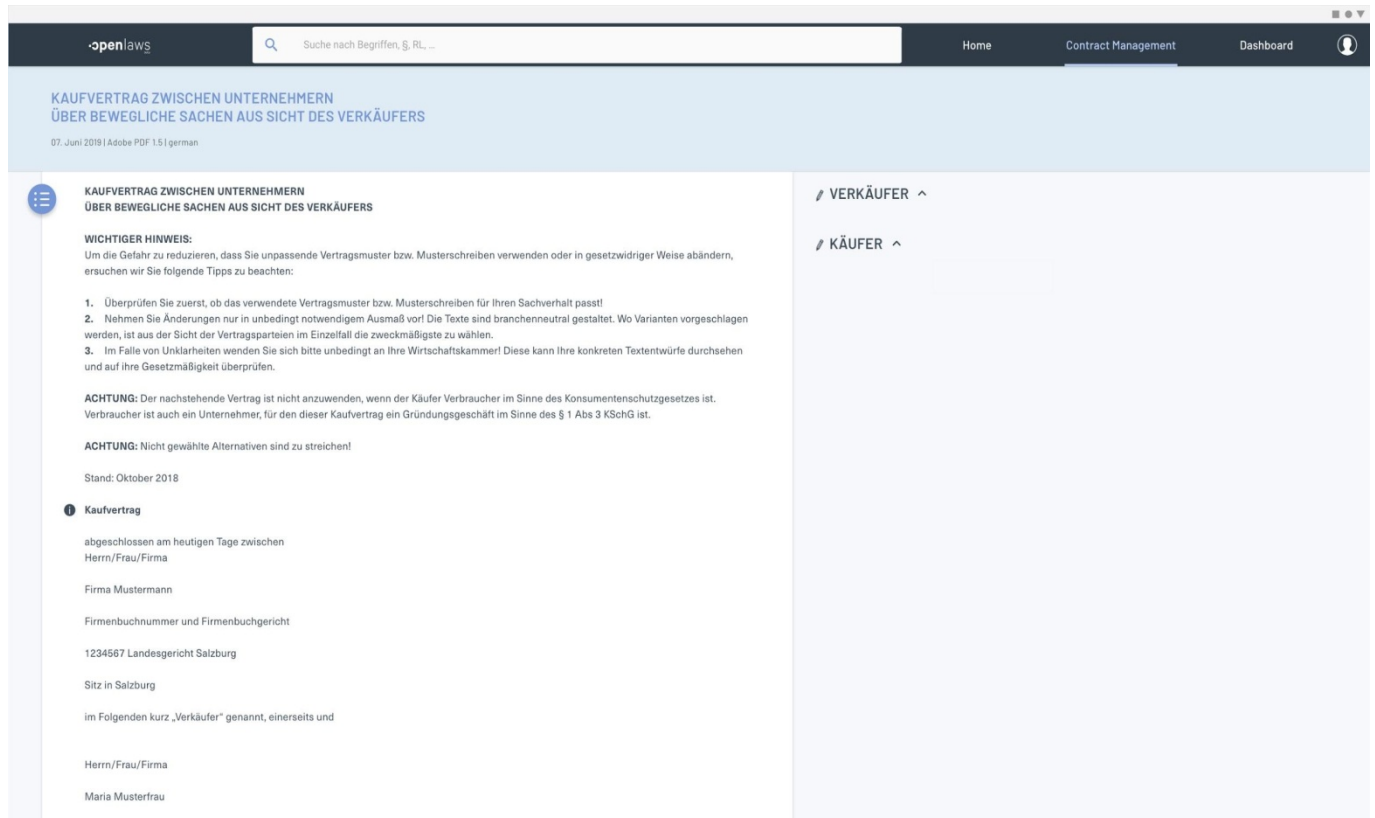


Figure 2. Main page for a single contract

Figure 3 shows the table of contents which is hidden behind the blue button on the left side. The user can click on a certain heading in order to instantly jump to the desired position.

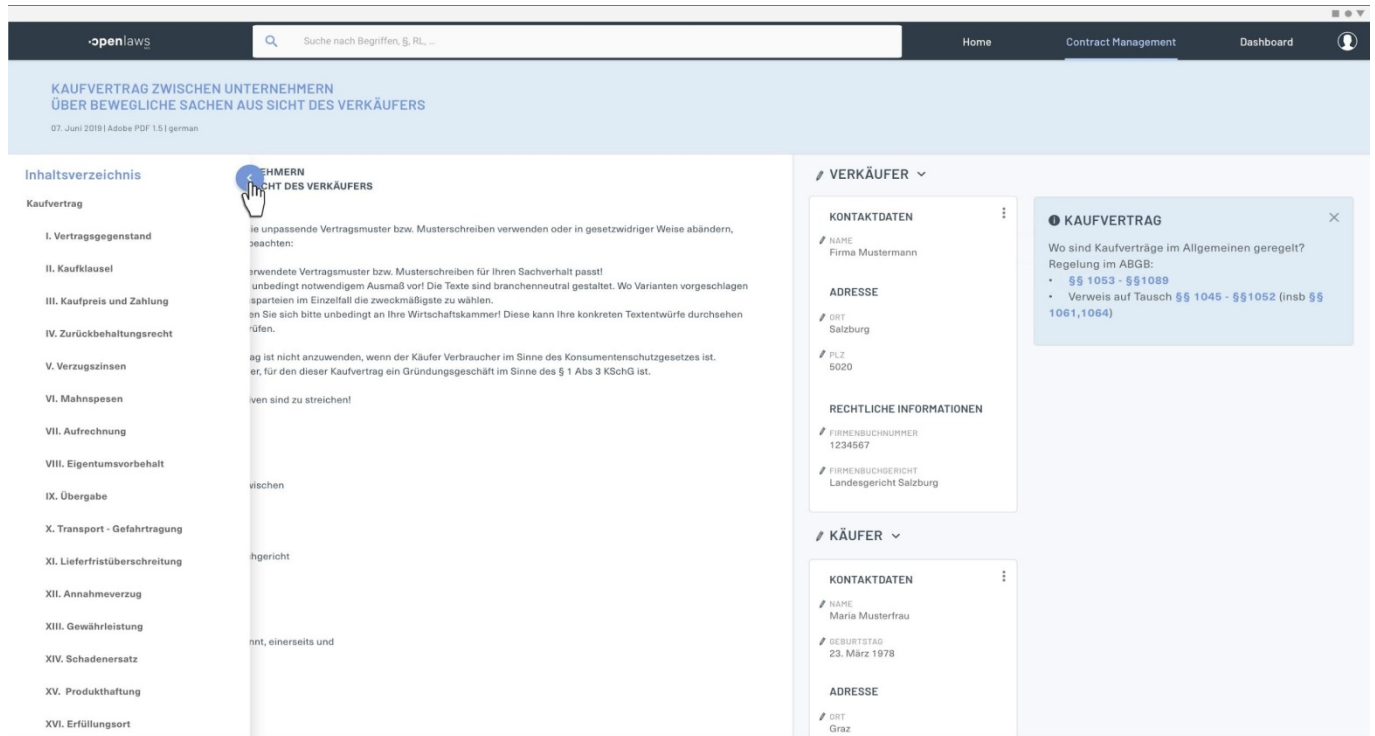


Figure 3. Table of Content

Figure 4 displays the detailed view of an extracted information unit. In this case it displays information about the seller ("Verkäufer"). The text passages from which we obtain the relevant information are highlighted in a light blue colour. These annotations are only visible when the respective detailed view is open.

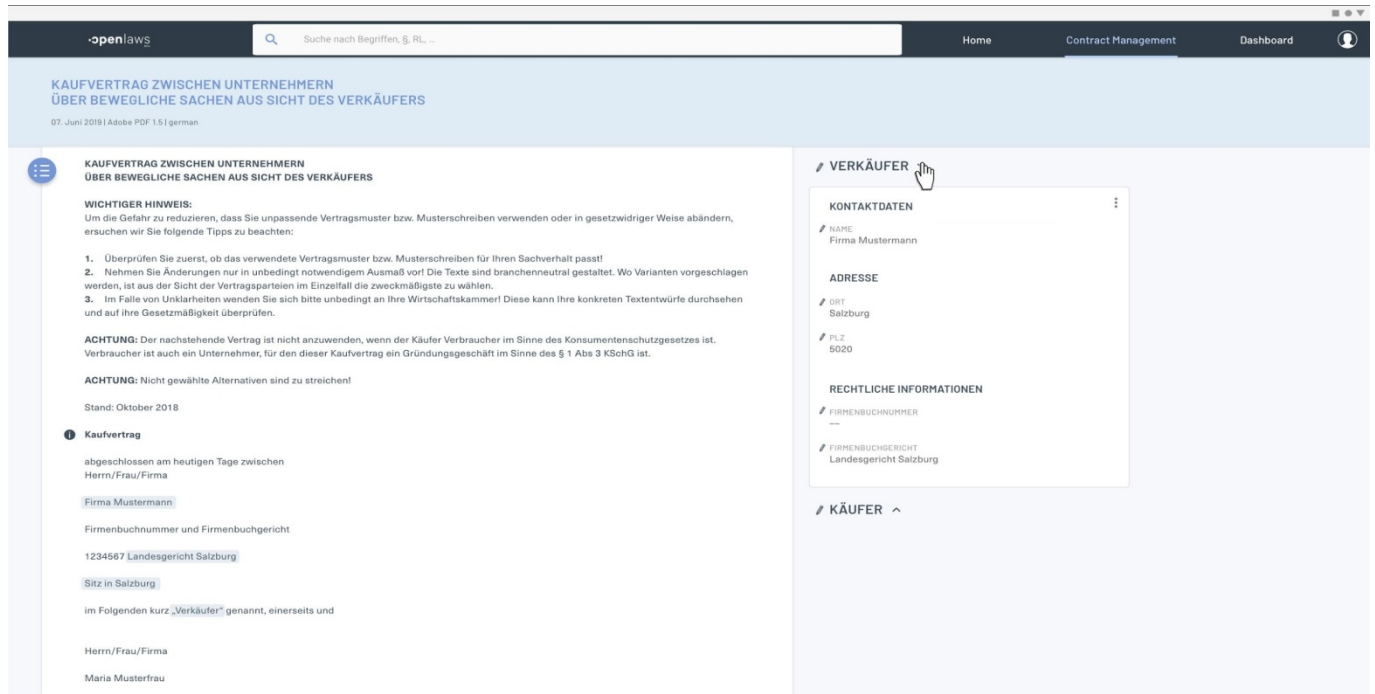
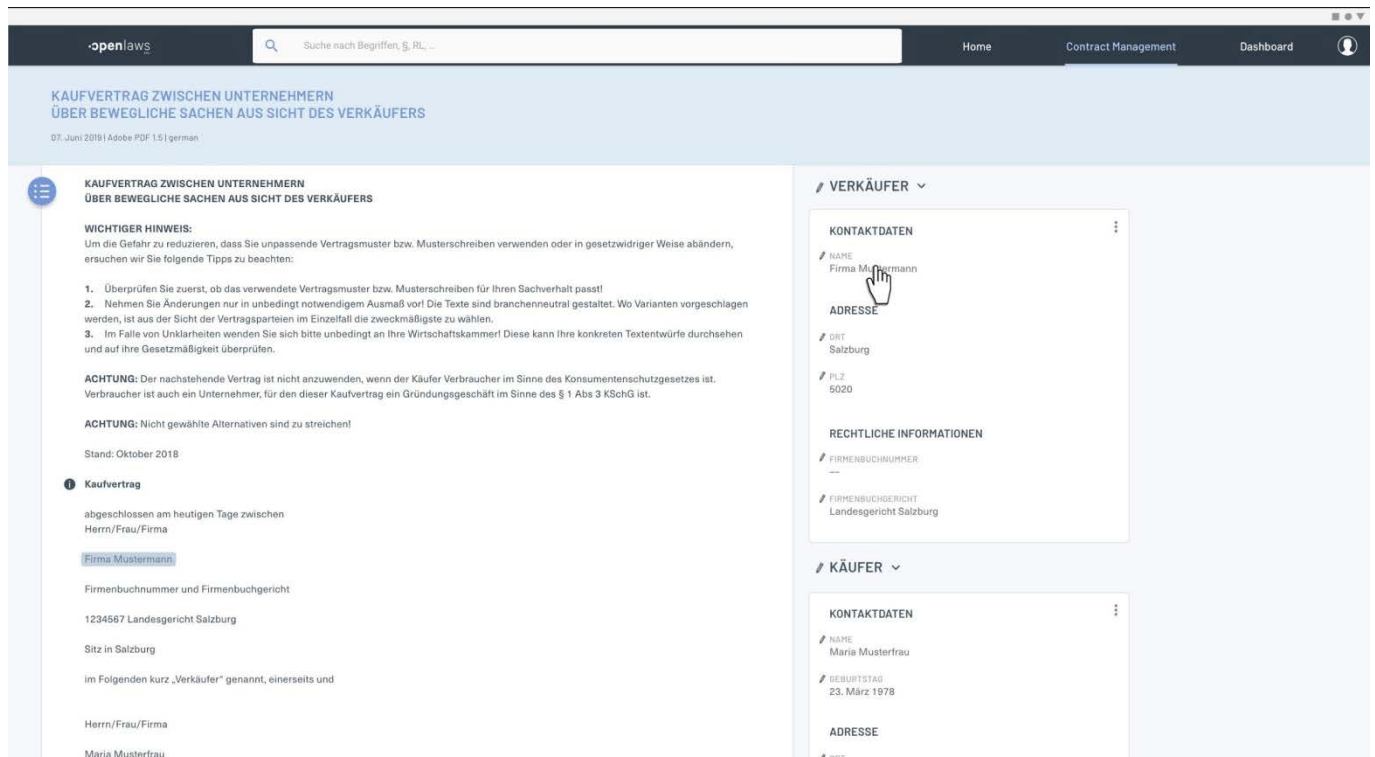


Figure 4. Detailed View of an Extracted Information Unit (Seller)

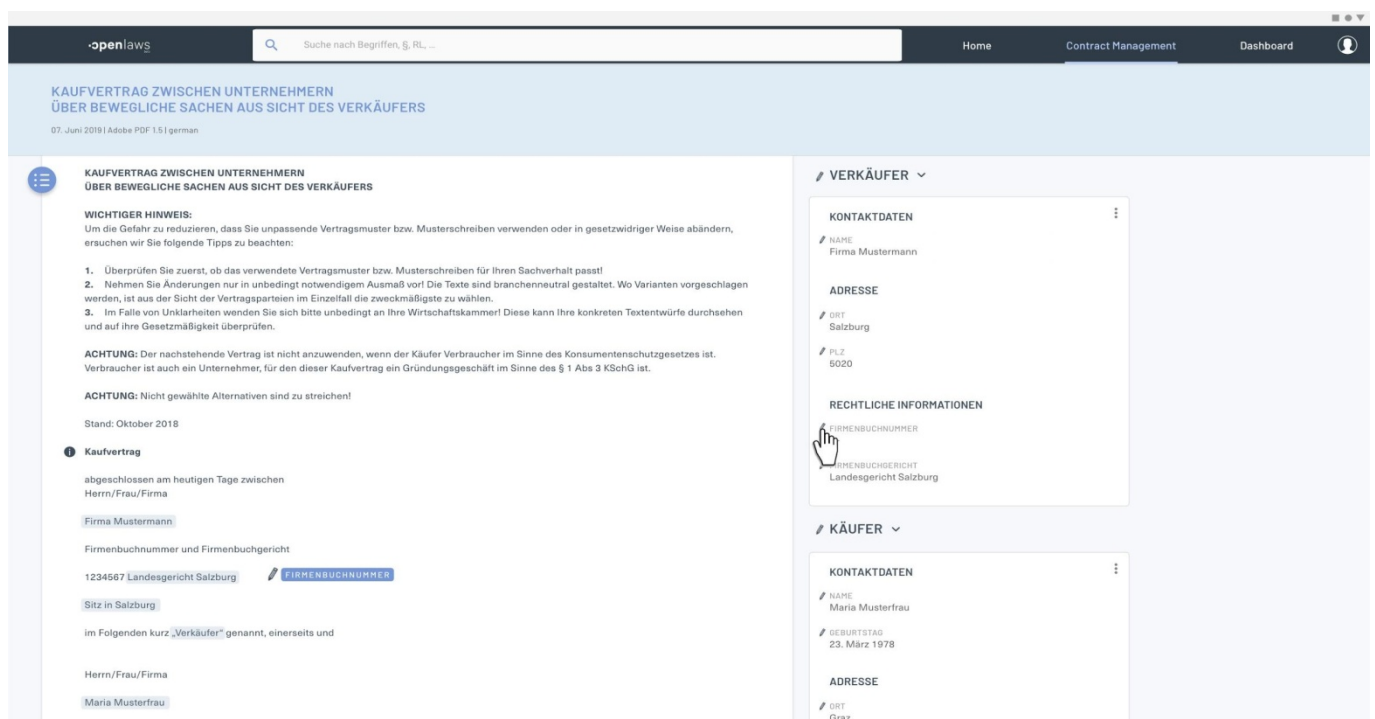
The user can select any of the annotations on the right-hand side to mark the position in the document. It helps users to find certain information in the document without reading the whole contract again. Figure 5 shows an example of this information query.



The screenshot shows the openlaws interface. The main document is titled "KAUFVERTRAG ZWISCHEN UNTERNEHMERN ÜBER BEWEGLICHE SACHEN AUS SICHT DES VERKÄUFERS". The sidebar on the right contains two sections: "VERKÄUFER" and "KÄUFER". Each section has a "KONTAKTDATEN" and "ADRESSE" unit. The "VERKÄUFER" section also has a "RECHTLICHE INFORMATIONEN" unit. The "KÄUFER" section has a "RECHTLICHE INFORMATIONEN" unit. The main document text includes a "WICHTIGER HINWEIS" section, a list of instructions, and a "Kaufvertrag" section. The sidebar units are populated with data from the document, such as "Firma Mustermann" for the seller and "Maria Musterfrau" for the buyer.

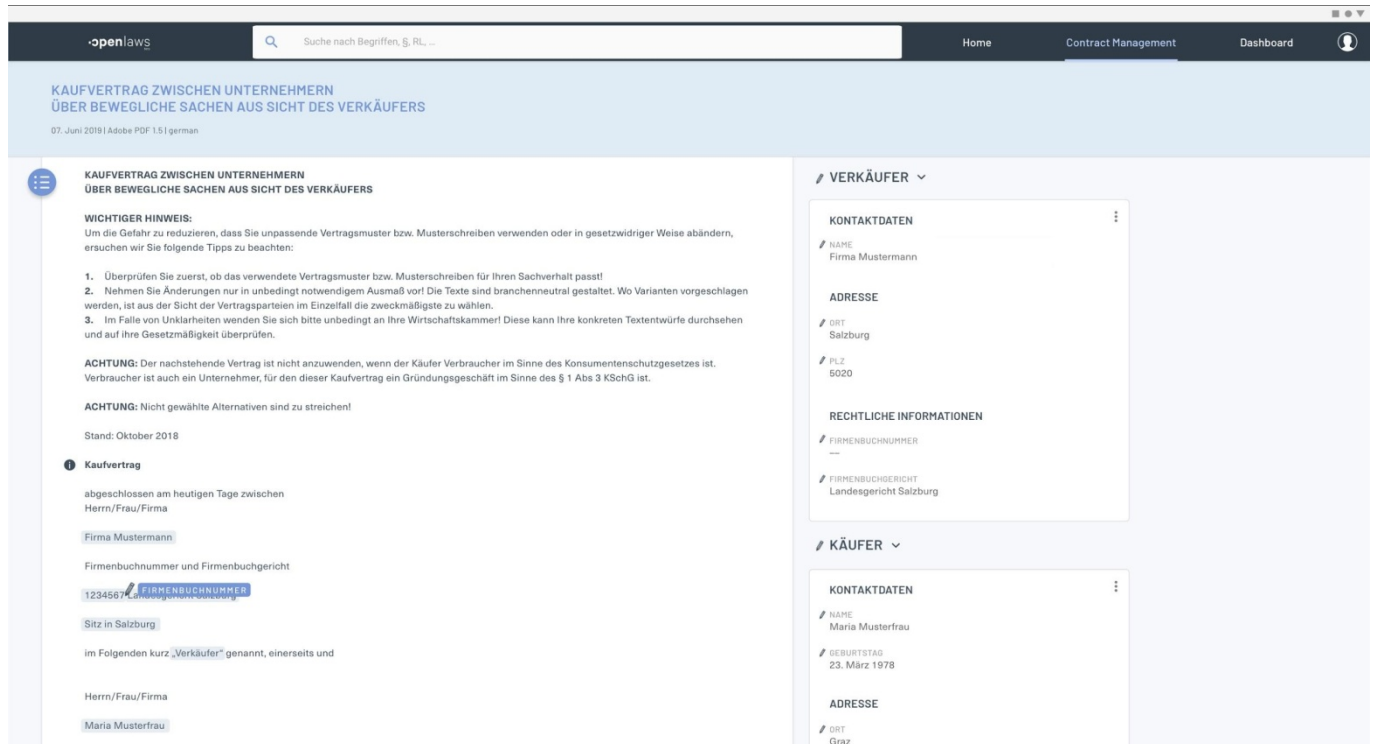
Figure 5. Information Query

If the system has not recognized the important information completely, the user can add it manually, preferably, by highlighting the relevant text passages. To do so, the user has to click on the pencil next to the title of the respective information unit. A mouse cursor to annotate the new information will appear. To get a new tag, the user has to mark the position in the document and drop the cursor. Afterwards, the new annotation appears on the right-hand side at the appropriate position. Figure 6, 7 and 8 illustrate this process.



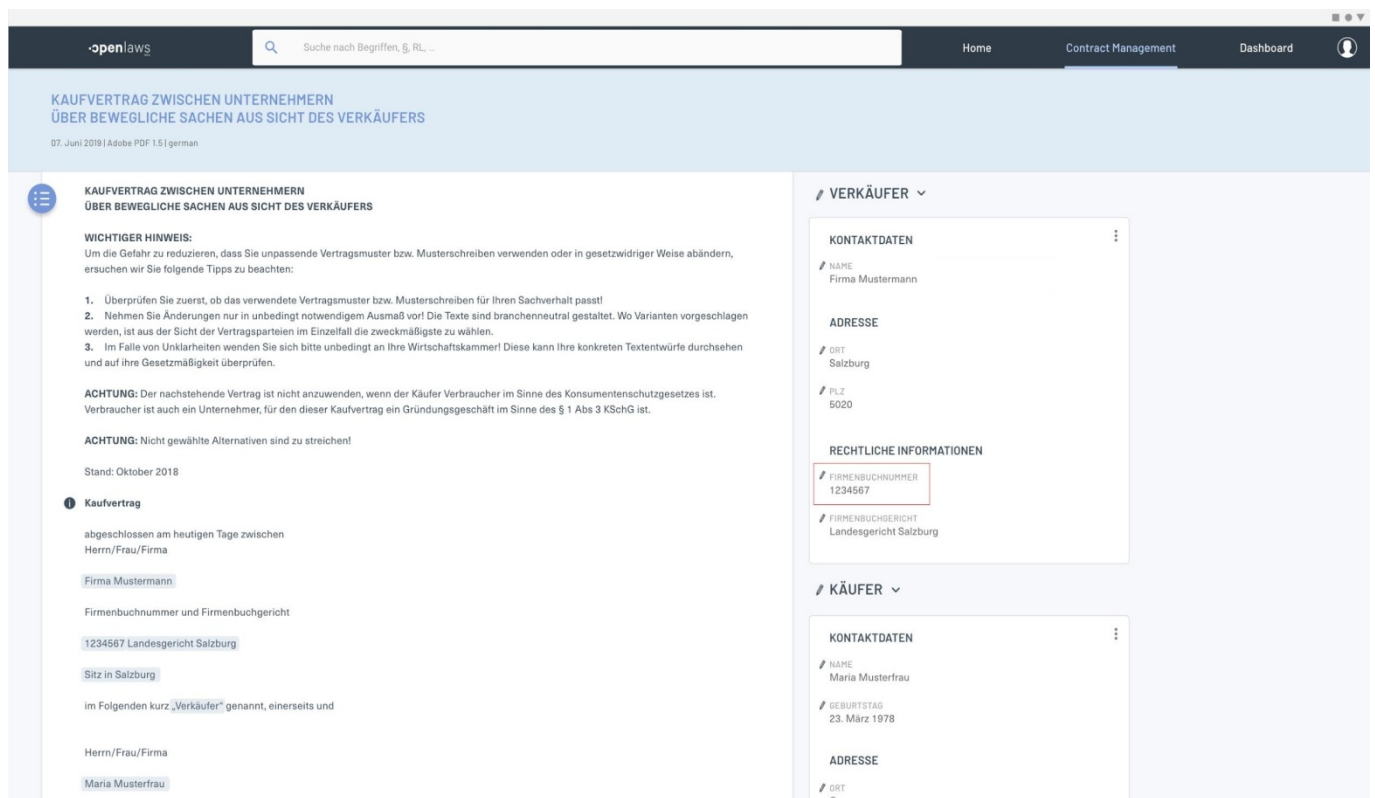
This screenshot shows the same interface as Figure 5, but with a mouse cursor hovering over the "FIRMBUCHNUMMER" unit in the "VERKÄUFER" section. The cursor is positioned over the unit title, indicating that the user is about to click on it to add new information. The main document text remains the same, but the sidebar units are updated with the new information added by the user.

Figure 6. Manually Add Information (Part 1)



The screenshot shows the openlaws interface. The main document is titled "KAUFVERTRAG ZWISCHEN UNTERNEHMERN ÜBER BEWEGLICHE SACHEN AUS SICHT DES VERKÄUFERS". The sidebar on the right is titled "VERKÄUFER" and contains two sections: "KONTAKTDATEN" and "RECHTLICHE INFORMATIONEN". The "KONTAKTDATEN" section includes fields for NAME (Firma Mustermann), ADRESSE (ORT: Salzburg, PLZ: 5020), and RECHTLICHE INFORMATIONEN (FIRMENBUCHNUMMER: --, FIRMENBUCHRICHT: Landesgericht Salzburg). The "KÄUFER" section is also visible, showing NAME (Maria Musterfrau), DEBUTSTAG (23. März 1978), and ADRESSE (ORT: Graz).

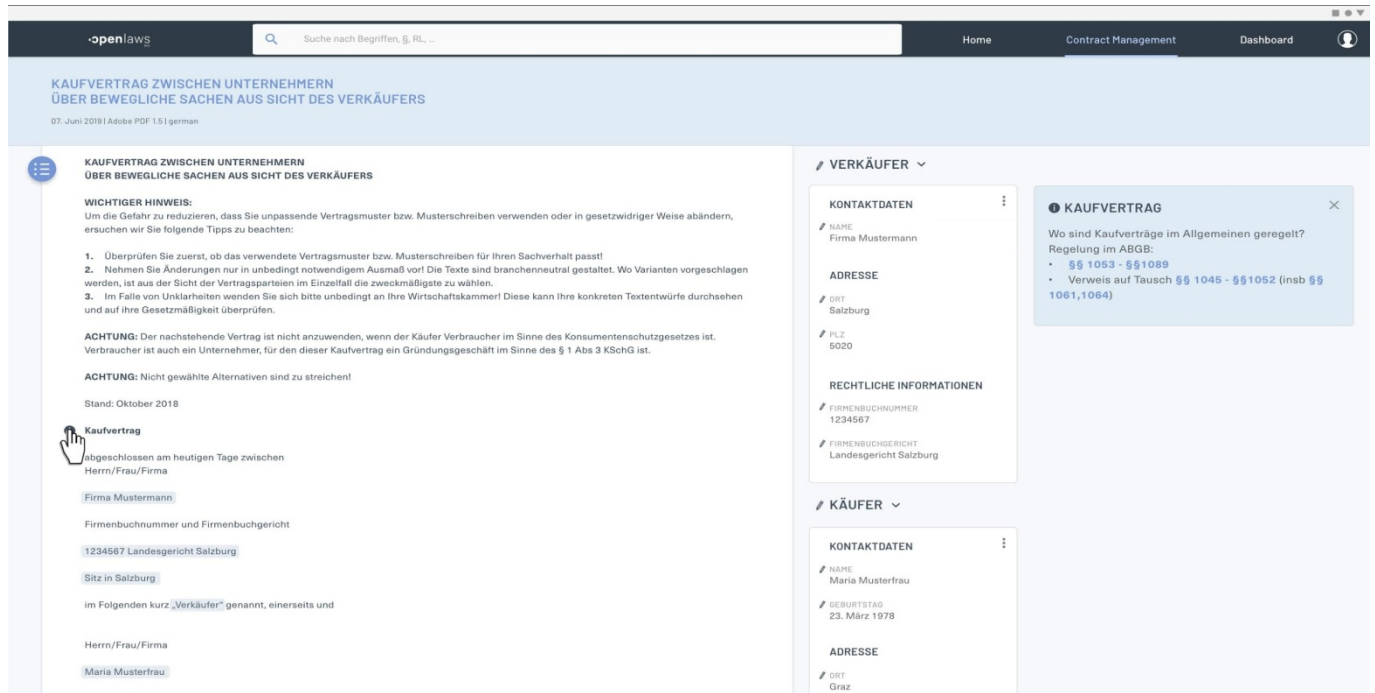
Figure 7. Manually Add Information (Part 2)



The screenshot shows the openlaws interface. The main document is titled "KAUFVERTRAG ZWISCHEN UNTERNEHMERN ÜBER BEWEGLICHE SACHEN AUS SICHT DES VERKÄUFERS". The sidebar on the right is titled "VERKÄUFER" and contains two sections: "KONTAKTDATEN" and "RECHTLICHE INFORMATIONEN". The "KONTAKTDATEN" section includes fields for NAME (Firma Mustermann), ADRESSE (ORT: Salzburg, PLZ: 5020), and RECHTLICHE INFORMATIONEN (FIRMENBUCHNUMMER: 1234567, FIRMENBUCHRICHT: Landesgericht Salzburg). The "KÄUFER" section is also visible, showing NAME (Maria Musterfrau), DEBUTSTAG (23. März 1978), and ADRESSE (ORT: Graz).

Figure 8. Manually Add Information (Part 3)

The user can also get further legal information about the contract. An info icon is shown next to every identified headline of the document. If the user presses the info icon, a box will appear on the far right (Figure 9). In the box the user gets information where e.g. the purchase agreement ("Kaufvertrag") is defined in the legislation and further references. This information is provided by the existing openlaws system based on the annotations of the document.



KAUFVERTRAG ZWISCHEN UNTERNEHMERN ÜBER BEWEGLICHE SACHEN AUS SICHT DES VERKÄUFERS

07. Juni 2018 | Adobe PDF 1.5 | german

WICHTIGER HINWEIS:
Um die Gefahr zu reduzieren, dass Sie unpassende Vertragsmuster bzw. Musterschreiben verwenden oder in gesetzwidriger Weise abändern, ersuchen wir Sie folgende Tipps zu beachten:

1. Überprüfen Sie zuerst, ob das verwendete Vertragsmuster bzw. Musterschreiben für Ihren Sachverhalt passt!
2. Nehmen Sie Änderungen nur in unbedingt notwendigem Ausmaß vor! Die Texte sind branchenneutral gestaltet. Wo Varianten vorgeschlagen werden, ist aus der Sicht der Vertragsparteien im Einzelfall die zweckmäßigste zu wählen.
3. Im Falle von Unklarheiten wenden Sie sich bitte unbedingt an Ihre Wirtschaftskammer! Diese kann Ihre konkreten Textentwürfe durchsehen und auf Ihre Gesetzmäßigkeit überprüfen.

ACHTUNG: Der nachstehende Vertrag ist nicht anzuwenden, wenn der Käufer Verbraucher im Sinne des Konsumentenschutzgesetzes ist. Verbraucher ist auch ein Unternehmer, für den dieser Kaufvertrag ein Gründungsgeschäft im Sinne des § 1 Abs 3 KSchG ist.

ACHTUNG: Nicht gewählte Alternativen sind zu streichen!

Stand: Oktober 2018

Kaufvertrag
abgeschlossen am heutigen Tage zwischen
Herrn/Frau/Firma

Firma Mustermann

Firmenbuchnummer und Firmenbuchgericht

1234567 Landesgericht Salzburg

Sitz in Salzburg

im Folgenden kurz „Verkäufer“ genannt, einerseits und

Herrn/Frau/Firma

Maria Musterfrau

VERKÄUFER

KONTAKTDATEN

NAME
Firma Mustermann

ADRESSE

ORT
Salzburg

PLZ
5020

RECHTLICHE INFORMATIONEN

FIRMENBUCHNUMMER
1234567

FIRMENBUCHGERICHT
Landesgericht Salzburg

KÄUFER

KONTAKTDATEN

NAME
Maria Musterfrau

GEBURTSTAG
23. März 1978

ADRESSE

ORT
Graz

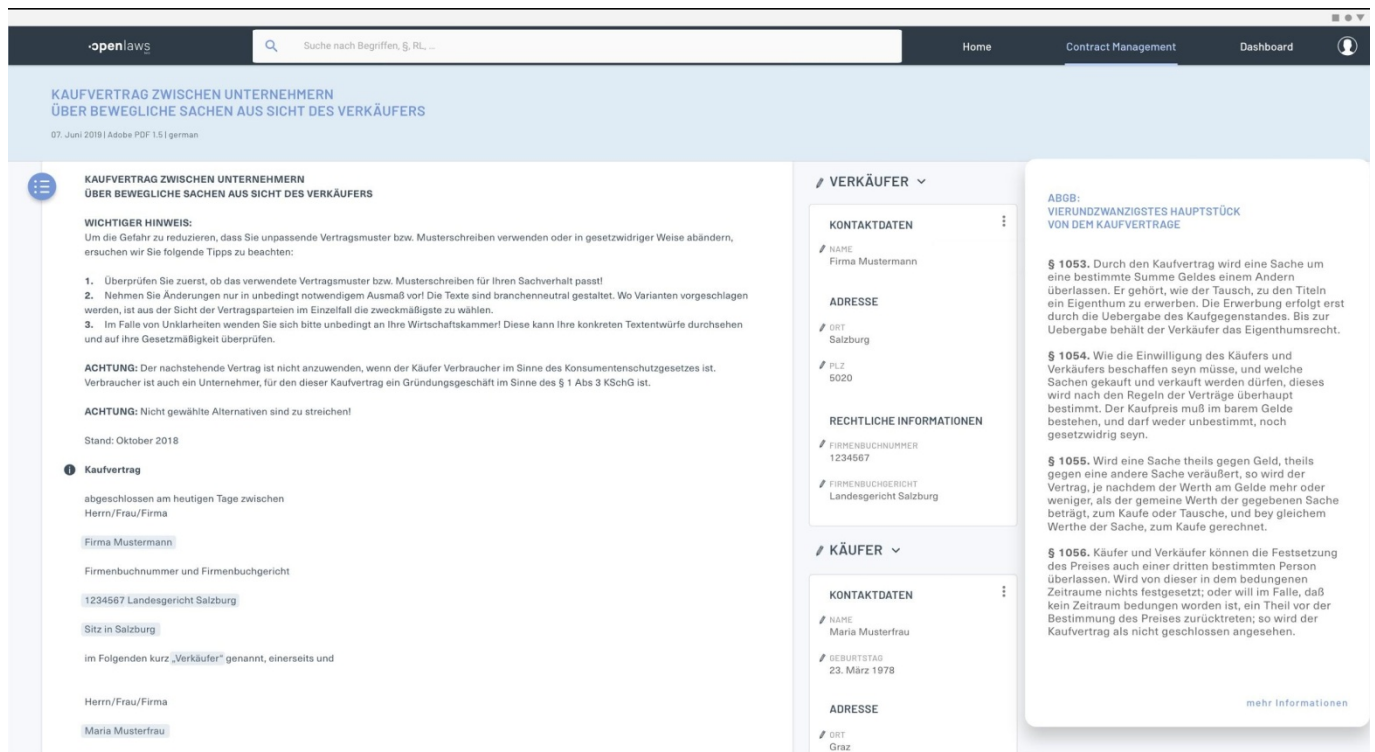
KAUFVERTRAG

Wo sind Kaufverträge im Allgemeinen geregelt?
Regelung im ABGB:

- §§ 1053 - §§ 1089
- Verweis auf Tausch §§ 1045 - §§ 1052 (insb §§ 1061, 1064)

Figure 9. Further Legal Information

The user can click on the related regulatory information, e.g. §§1053 - §§1089 to find out more about it within the legislation without leaving the page (Figure 10).



KAUFVERTRAG ZWISCHEN UNTERNEHMERN ÜBER BEWEGLICHE SACHEN AUS SICHT DES VERKÄUFERS

07. Juni 2018 | Adobe PDF 1.5 | german

WICHTIGER HINWEIS:
Um die Gefahr zu reduzieren, dass Sie unpassende Vertragsmuster bzw. Musterschreiben verwenden oder in gesetzwidriger Weise abändern, ersuchen wir Sie folgende Tipps zu beachten:

1. Überprüfen Sie zuerst, ob das verwendete Vertragsmuster bzw. Musterschreiben für Ihren Sachverhalt passt!
2. Nehmen Sie Änderungen nur in unbedingt notwendigem Ausmaß vor! Die Texte sind branchenneutral gestaltet. Wo Varianten vorgeschlagen werden, ist aus der Sicht der Vertragsparteien im Einzelfall die zweckmäßigste zu wählen.
3. Im Falle von Unklarheiten wenden Sie sich bitte unbedingt an Ihre Wirtschaftskammer! Diese kann Ihre konkreten Textentwürfe durchsehen und auf Ihre Gesetzmäßigkeit überprüfen.

ACHTUNG: Der nachstehende Vertrag ist nicht anzuwenden, wenn der Käufer Verbraucher im Sinne des Konsumentenschutzgesetzes ist. Verbraucher ist auch ein Unternehmer, für den dieser Kaufvertrag ein Gründungsgeschäft im Sinne des § 1 Abs 3 KSchG ist.

ACHTUNG: Nicht gewählte Alternativen sind zu streichen!

Stand: Oktober 2018

Kaufvertrag
abgeschlossen am heutigen Tage zwischen
Herrn/Frau/Firma

Firma Mustermann

Firmenbuchnummer und Firmenbuchgericht

1234567 Landesgericht Salzburg

Sitz in Salzburg

im Folgenden kurz „Verkäufer“ genannt, einerseits und

Herrn/Frau/Firma

Maria Musterfrau

VERKÄUFER

KONTAKTDATEN

NAME
Firma Mustermann

ADRESSE

ORT
Salzburg

PLZ
5020

RECHTLICHE INFORMATIONEN

FIRMENBUCHNUMMER
1234567

FIRMENBUCHGERICHT
Landesgericht Salzburg

KÄUFER

KONTAKTDATEN

NAME
Maria Musterfrau

GEBURTSTAG
23. März 1978

ADRESSE

ORT
Graz

ABGB: VIERUNDZWANZIGSTES HAUPTSTÜCK VON DEM KAUFVERTRAGE

§ 1053. Durch den Kaufvertrag wird eine Sache um eine bestimmte Summe Geldes einem Andern überlassen. Er gehört, wie der Tausch, zu den Titeln ein Eigentum zu erwerben. Die Erwerbung erfolgt erst durch die Uebergabe des Kaufgegenstandes. Bis zur Uebergabe behält der Verkäufer das Eigentumsrecht.

§ 1054. Wie die Einwilligung des Käufers und Verkäufers beschaffen seyn müsse, und welche Sachen gekauft und verkauft werden dürfen, dieses wird nach den Regeln der Verträge überhaupt bestimmt. Der Kaufpreis muß im barem Gelde bestehen, und darf weder unbestimmt, noch gesetzwidrig seyn.

§ 1055. Wird eine Sache theils gegen Geld, theils gegen eine andere Sache veräußert, so wird der Vertrag, je nachdem der Werth am Gelde mehr oder weniger, als der gemeine Werth der gegebenen Sache beträgt, zum Kaufe oder Tausche, und bey gleichem Werthe der Sache, zum Kaufe gerechnet.

§ 1056. Käufer und Verkäufer können die Festsetzung des Preises auch einer dritten bestimmten Person überlassen. Wird von dieser in dem bedungenen Zeitraume nichts festgesetzt; oder will im Falle, daß kein Zeitraum bedungen worden ist, ein Theil vor der Bestimmung des Preises zurücktreten; so wird der Kaufvertrag als nicht geschlossen angesehen.

mehr Informationen

Figure 10. Further Legal Information (Detailed View)

2.5 CURRENT STAGE OF THE PILOT

Currently our main focus is to improve quality in the results of annotations and the type of annotations, this is obviously a critical point in providing a quality customer experience and to differentiating with a pure search solution. The following tasks will be performed until the end of the project:

- train the system on additional fine-grained NER types not only Company, Country and Location, to improve the quality of the named entities.
- train the RelEx on new relations, as the existing ones are not sufficient for the recognition of “Buyer”, “Seller” or other parties.
- retrain language models
- improve usability based on the feedback,
- stabilize the system,
- add documentation (installation, user)

2.6 WORKING BASIS IN THE PILOT

To proof our development, we had different document sets of real customers of the Cybly GmbH (former openlaws). In the beginning we tried to work with publicly available contract templates, which proved to be inefficient, as general templates did not provide the variety needed. We had a rich dataset of approx. 100 000 documents (of different type, e.g. contracts, communication and format, e.g. PDF, MS-word, email) which we used to optimize the conversion process from the different formats to a Lynx-Documents. Furthermore, we extracted additional information, such as document dates, customer IDs, offer numbers, etc. We aimed also identifying **index clauses**, that are those clauses that refer to the price being tied to a consumer price index to ensure value stability (Figure 11 **Sample of an index clause**). With the ability to determine index clauses, we enable the identification of **other contractual clauses** too.

“Index Assurance: Prices are subject to annual value adjustment in accordance with the Consumer Price Index (CPI) 2010 or the index replacing it. The starting point for the calculation of value assurance is the index number published for the month in which the offer was placed.”

Figure 11 Sample of an index clause

Working with this real data was very beneficial in terms of testing the capability and the performance of the application. However, this means working with highly confidential information. For this reason, we agreed with our customer on performing all tests within the customer’s environment, where we deploy parts of the Cybly development and the Lynx System. Unfortunately, we were not able to go ahead with this dataset as our customer was not allowed to continue working on this project due to the COVID-19 crisis.

In addition to the first data set mentioned above, we use a testing data set of approximately 10 000 documents, which provides a collection of different contract types, e.g. non-disclosure agreements, rental contracts, decisions of civil authorities, etc. This dataset offers a wide variety, as the contracts originate from different sources. This data set is already well-structured. The documents are grouped by contract partner, type, etc.

Finally, we have a dataset of approx. 2TB of different types and format including mailboxes which will be used in the evaluation phase to answer real questions of our partner company. The ingestion of this document has not yet begun.

3 PILOT DESCRIPTION

This section describes the solution and the visualizations of the front end based on the main objectives from section 2.

3.1 HARVESTING

To harvest documents a command line tool has been implemented which provides the following two main functionalities:

- Recursively send all documents of a directory and its subdirectory to the Document Service for processing them
- Monitoring a given directory and its subdirectory and send notifications when the contents of the specified files or directories are modified

With this tool it is possible to ingest a large set of documents to the system and also to monitor this set for any subsequent changes. This tool can be installed as a service on Windows, Linux and MacOS.

3.2 CONVERSION

The Conversion Service is a stand-alone application which provides the functionality to convert different document formats to a Lynx-Document. In general, this is a pure back end service without any user interface, but for demonstration and promotion purpose a small UI has been developed where documents in different formats can be uploaded and converted. The converted document isn't represented as Lynx-Document itself, but all of its information.

The following main document formats are currently supported:

Name of the format	Remark
Microsoft Office document formats	Microsoft Office 97 onwards, Word Excel and Powerpoint are currently filtered but can be enabled on request
Open Document Format	
iWorks	Pages, Number and Keynote are currently filtered
HTML	
PDF	Structure is per page, incl. OCR
Images	As input for the OCR
Outlook Messages (*.msg)	Including attachment, the convertert returns a list of document
MIME Messages	Including attachment, the convertert returns a list of document

Table 1. Supported document formats

Figure 12 shows the upload page where a user can drop a document and request the conversion. Figure 13 shows the converted document. This demo makes also use of the Extraction / Annotation Service

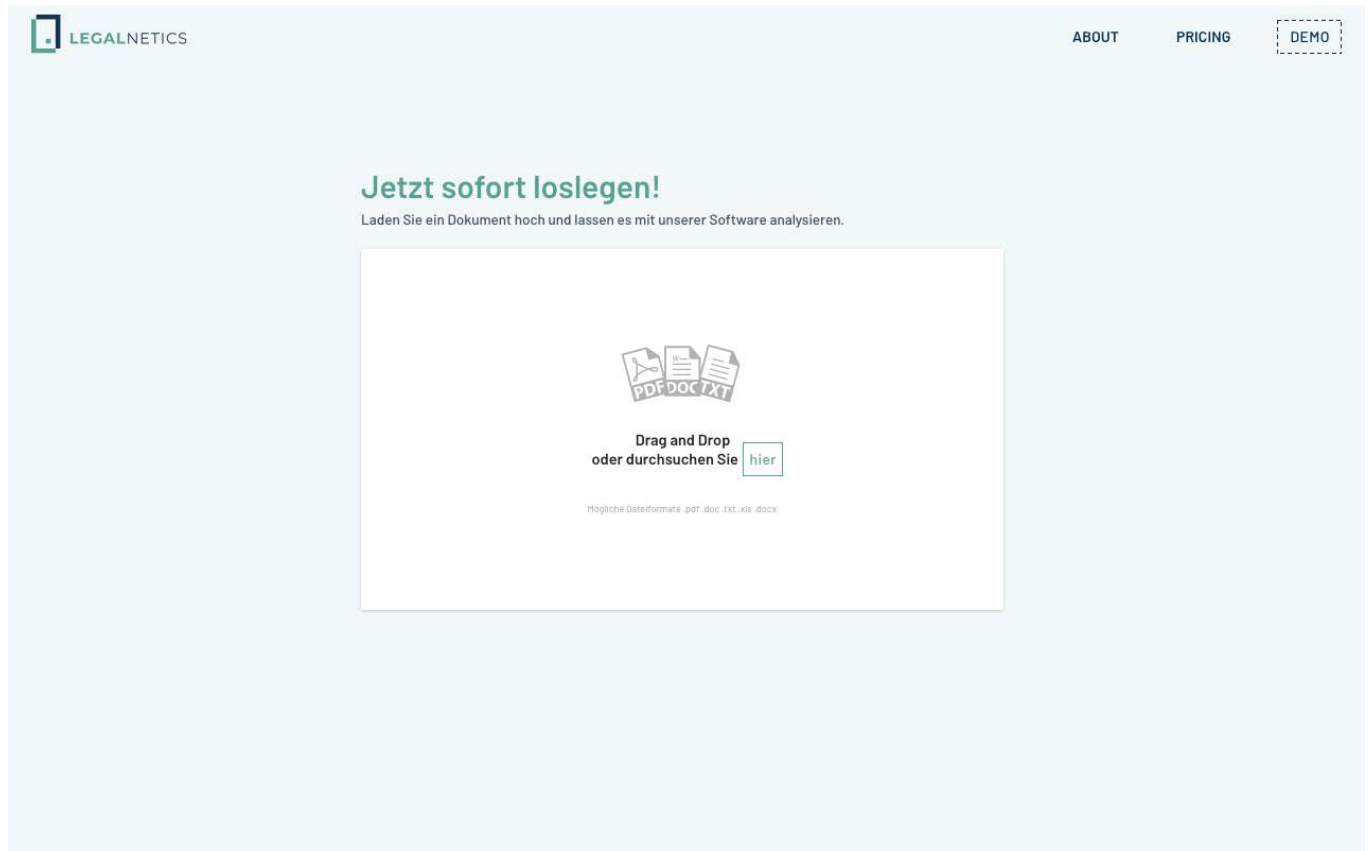


Figure 12 Demo Document-Upload

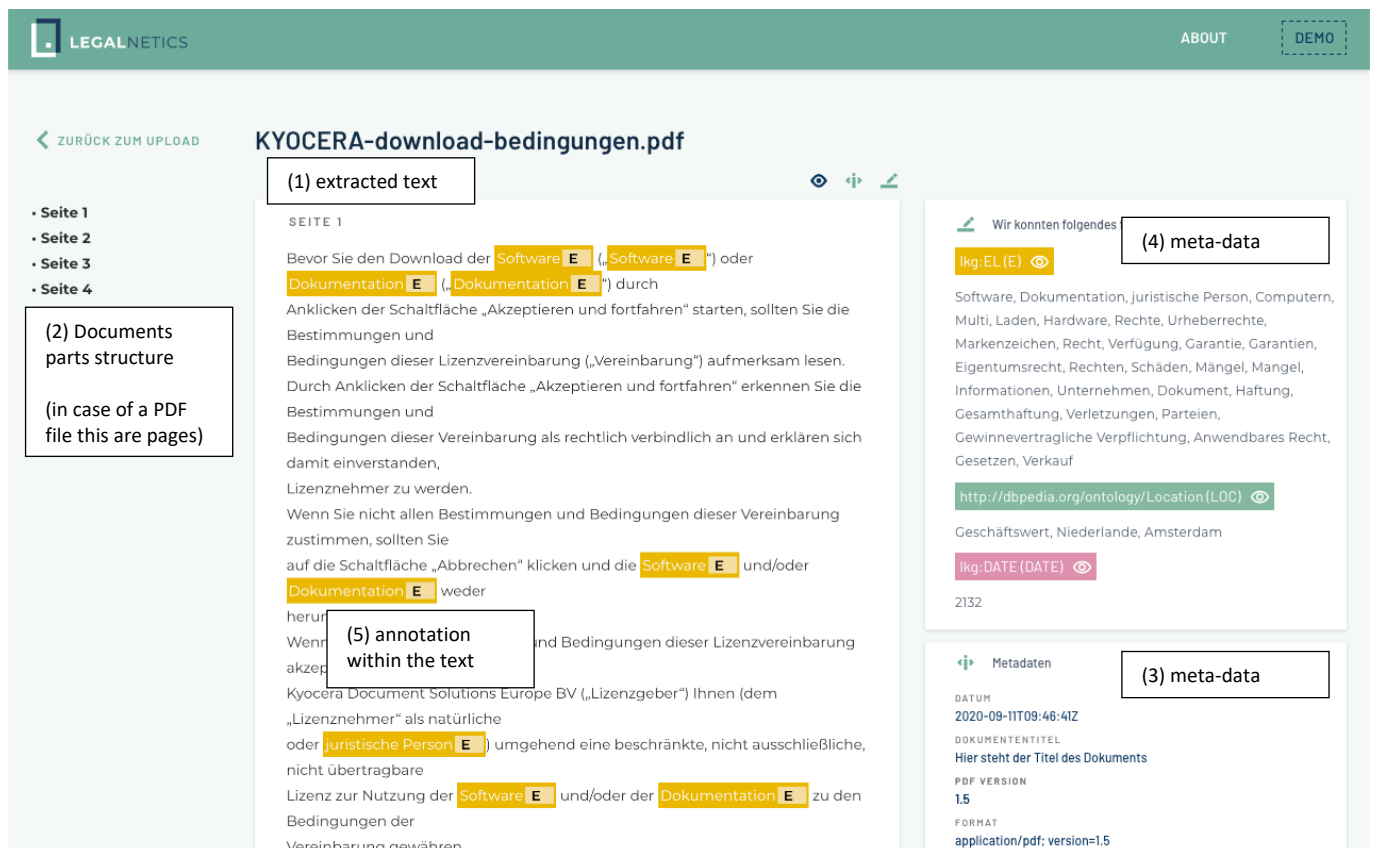


Figure 13 Demo Document-View

In the center the extracted text (1) is shown, on the left-hand side the document structure which is based on parts (2). On the right-hand side the meta-data (3) and extracted annotations (4) are shown. Each single annotation is also presented within the text (5)

3.3 EXTRACTION

The extraction is done by the Annotation Service which orchestrates the calls to the different Lynx Annotations Services and also to others. Main Annotation Services connected are:

- **TimEx** - Temporal extraction. To detect temporal entities within the document, e.g., dates when the offer has been made, durations, period of validity.
- **NER** - Named Entity Recognition – Is used to identify named entities such as persons and organisations (companies), courts, roles, administrations, group of persons, etc.
- **EntEx+WSID** - Entity Extraction with Word Sens Disambiguation Service – to enrich the documents with entities from the defined vocabulary created for this use case.
- **RelEx** - Relation Extraction between entities within a single document will be used to find e.g. cause-effect relation: Example: The agreement ends by 20.1.2020
- **Geo** – to find geographical expressions in documents, mainly addresses.
- **RegEx** – to find custom defined patterns within the document, e.g. order numbers, customer numbers, VAT numbers, company registration numbers, references to legislation, etc.

3.4 ARCHIVING

The document and its extracted information are stored within the LawThek² Document Store. LawThek is a core product of Cybly, it is a free legal information system that offers cross-platform access to standardized and interlinked legal information for multiple jurisdictions. LawThek currently provides access to European, Austrian and German legislation, jurisdiction and topic-based summaries. To do so, LawThek has been extended by the possibility to store Lynx-Documents with the annotations beside the original document.

Documents added by the harvester are added to the LawThek Document Store, but not directly visible through browsing the data store, in fact in a first step they can only be found through search. The reason for this is that allowing this would duplicate the structure which is already available through the file system and would not add benefits.

3.5 SEARCH

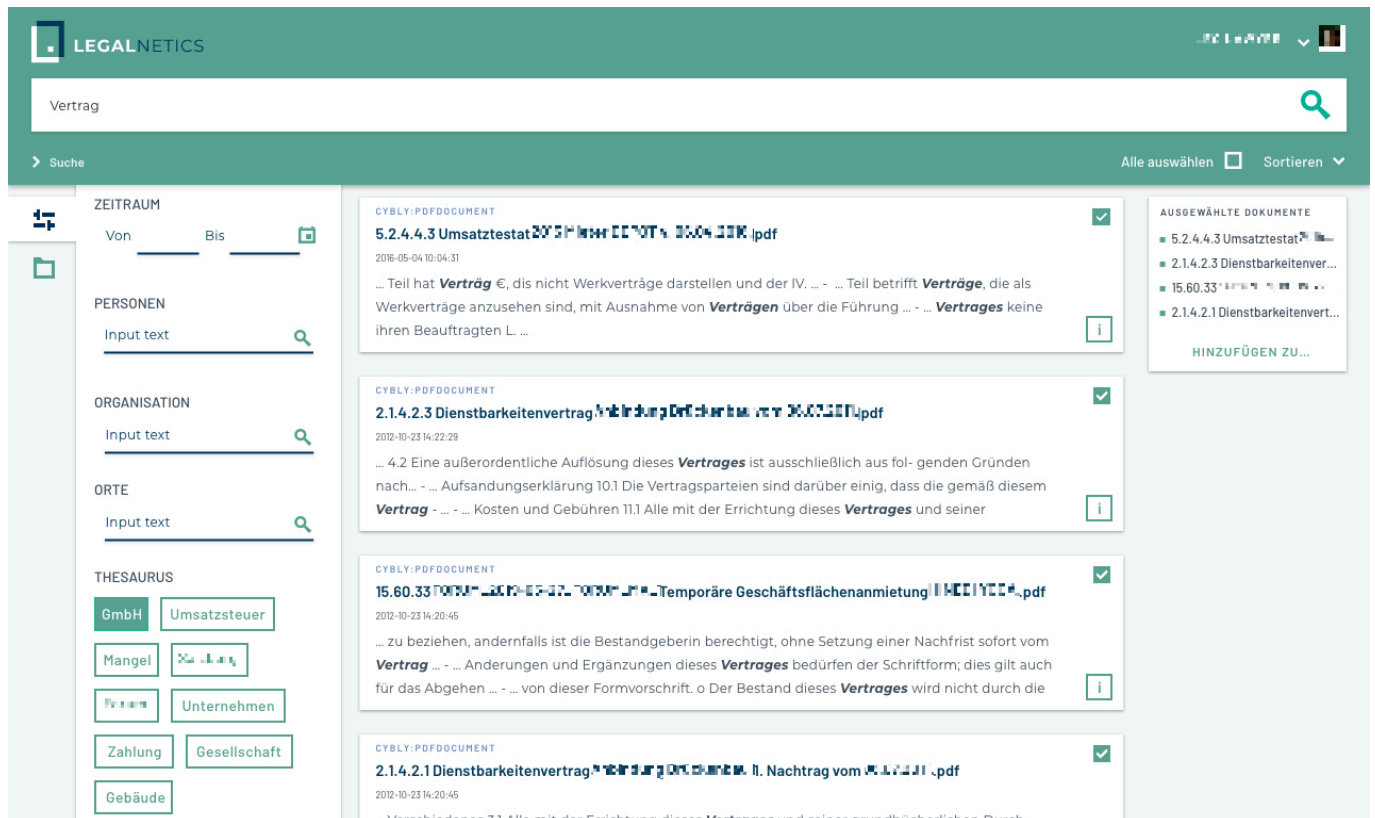
The ability to search for documents is a central feature of the newly developed solution. As for newly added documents, search is the only entry point to these documents.

The search builds on top of the Lynx SEAR service. It is possible to search for documents by full text, document type, annotations, metadata, e.g. document date (facets) and any combination. It is therefore possible to search for newly added / modified documents. Or for certain document names, etc.

The new search service replaces the existing search service within the LawThek. Figure 14 shows the new search result screen. On the left there are the different filters / facets based on the search result.

The found documents can be viewed or added to new / existing virtual folders (see 3.6 Manage).

² <https://lawthek.eu>



The screenshot shows the LEGALNETICS search interface. The search bar at the top contains the word "Vertrag". Below the search bar, there are filters for "ZEITRAUM" (Time Period), "PERSONEN" (Persons), "ORGANISATION" (Organisation), "ORTE" (Locations), and "THESAURUS" (Thesaurus). The search results are displayed in a list format, showing document titles, dates, and snippets of text. The results are filtered by "Vertrag" and "Verträge".

Search Results:

- 5.2.4.4.3 Umsatztestat** (2016-05-04 10:04:31)

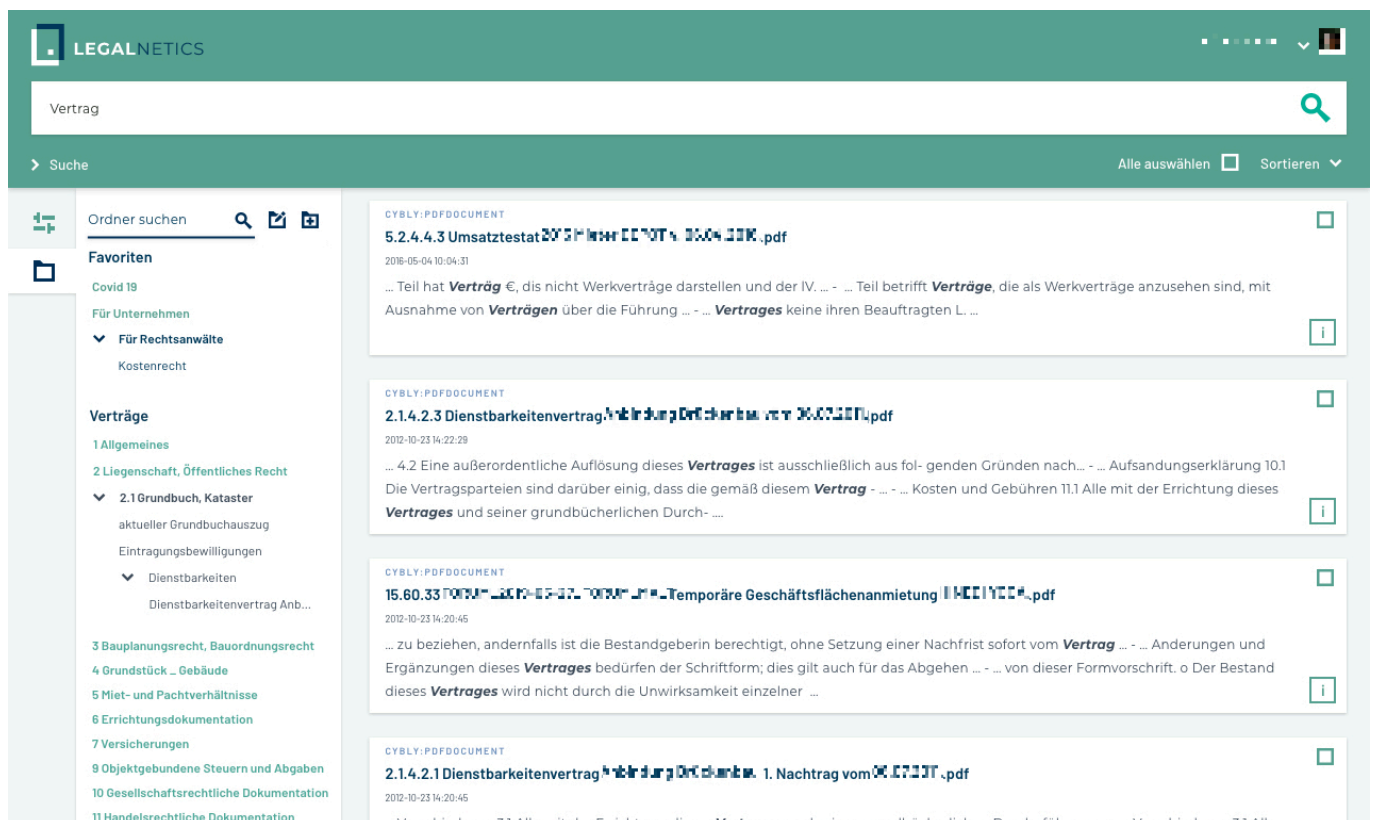
... Teil hat **Vertrag** €, dis nicht Werkverträge darstellen und der IV. ... Teil betrifft **Verträge**, die als Werkverträge anzusehen sind, mit Ausnahme von **Verträgen** über die Führung ... - ... **Vertrages** keine ihren Beauftragten L ...
- 2.1.4.2.3 Dienstbarkeitenvertrag** (2012-10-23 14:22:29)

... 4.2 Eine außerordentliche Auflösung dieses **Vertrages** ist ausschließlich aus fol- genden Gründen nach... - ... Aufsandungserklärung 10.1 Die Vertragsparteien sind darüber einig, dass die gemäß diesem **Vertrag** - ... - ... Kosten und Gebühren 11.1 Alle mit der Errichtung dieses **Vertrages** und seiner
- 15.60.33** (2012-10-23 14:20:45)

... zu beziehen, andernfalls ist die Bestandgeberin berechtigt, ohne Setzung einer Nachfrist sofort vom **Vertrag** ... - ... Änderungen und Ergänzungen dieses **Vertrages** bedürfen der Schriftform; dies gilt auch für das Abgehen ... - ... von dieser Formvorschrift. o Der Bestand dieses **Vertrages** wird nicht durch die
- 2.1.4.2.1 Dienstbarkeitenvertrag** (2012-10-23 14:20:45)

... Verschiedenes 3.1 Alle mit der Errichtuna dieses **Vertraaes** und seiner arundbücherlichen Durch-

Figure 14 Search with search result



The screenshot shows the LEGALNETICS search interface with a folder view. The search bar at the top contains the word "Vertrag". Below the search bar, there are filters for "Ordner suchen" (Search folders), "Favoriten" (Favorites), and "Verträge" (Contracts). The search results are displayed in a list format, showing document titles, dates, and snippets of text. The results are filtered by "Vertrag" and "Verträge".

Search Results:

- 5.2.4.4.3 Umsatztestat** (2016-05-04 10:04:31)

... Teil hat **Vertrag** €, dis nicht Werkverträge darstellen und der IV. ... Teil betrifft **Verträge**, die als Werkverträge anzusehen sind, mit Ausnahme von **Verträgen** über die Führung ... - ... **Vertrages** keine ihren Beauftragten L ...
- 2.1.4.2.3 Dienstbarkeitenvertrag** (2012-10-23 14:22:29)

... 4.2 Eine außerordentliche Auflösung dieses **Vertrages** ist ausschließlich aus fol- genden Gründen nach... - ... Aufsandungserklärung 10.1 Die Vertragsparteien sind darüber einig, dass die gemäß diesem **Vertrag** - ... - ... Kosten und Gebühren 11.1 Alle mit der Errichtung dieses **Vertrages** und seiner grundbücherlichen Durch- ...
- 15.60.33** (2012-10-23 14:20:45)

... zu beziehen, andernfalls ist die Bestandgeberin berechtigt, ohne Setzung einer Nachfrist sofort vom **Vertrag** ... - ... Änderungen und Ergänzungen dieses **Vertrages** bedürfen der Schriftform; dies gilt auch für das Abgehen ... - ... von dieser Formvorschrift. o Der Bestand dieses **Vertrages** wird nicht durch die Unwirksamkeit einzelner ...
- 2.1.4.2.1 Dienstbarkeitenvertrag** (2012-10-23 14:20:45)

... Verschiedenes 3.1 Alle mit der Errichtuna dieses **Vertraaes** und seiner arundbücherlichen Durch- führung ... - ... Verschiedenes 3.1 Alle

Figure 15 Search result, with folder view

3.6 MANAGE

LawThek offers the possibility to manage folders. A folder within LawThek is a group of different documents and folders. Folders can be private or shared. For shared folders the access rights can be defined based on individuals or access group level.

Individual items of the search result can be selected and moved to an existing folder or added to a new folder. An item can belong to multiple folders at the same time

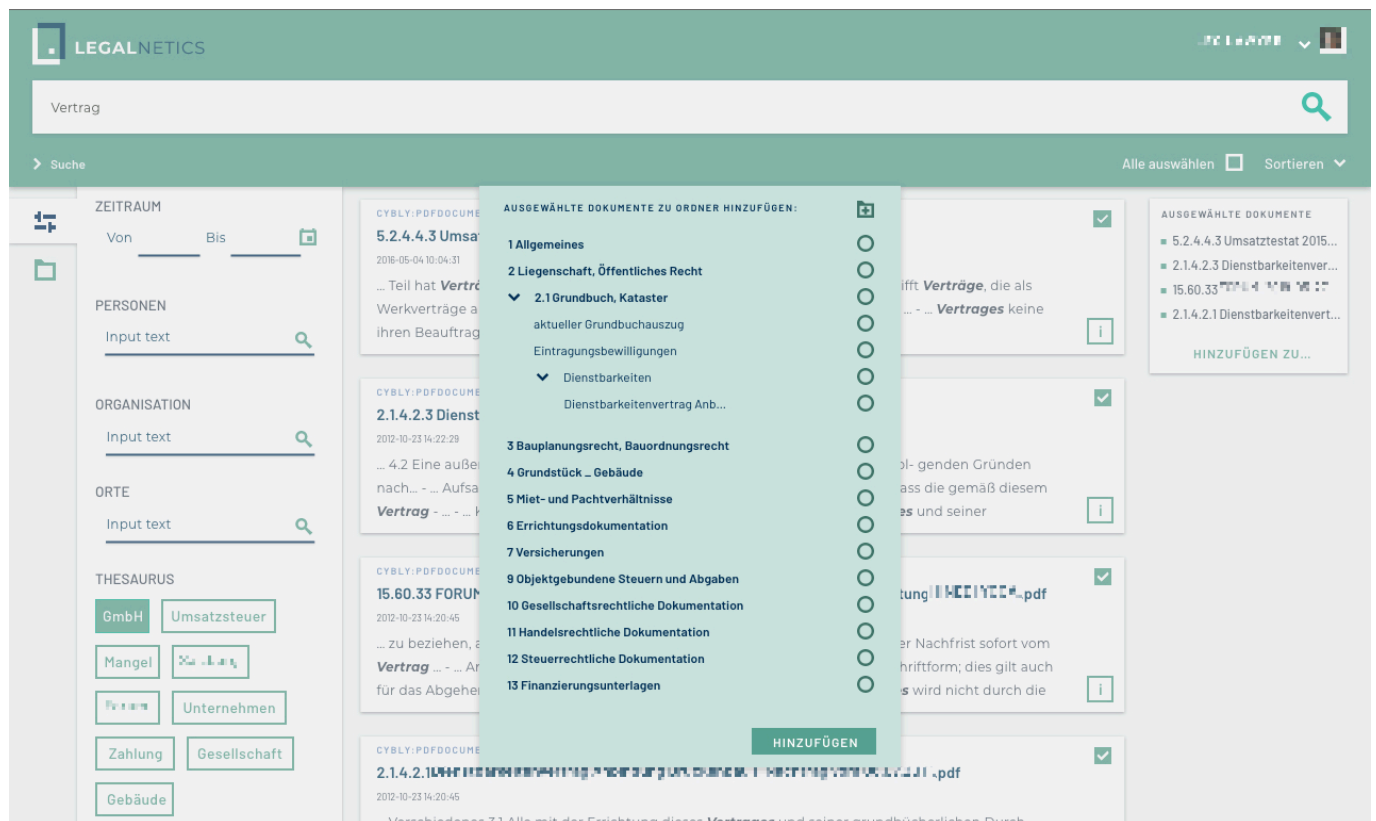
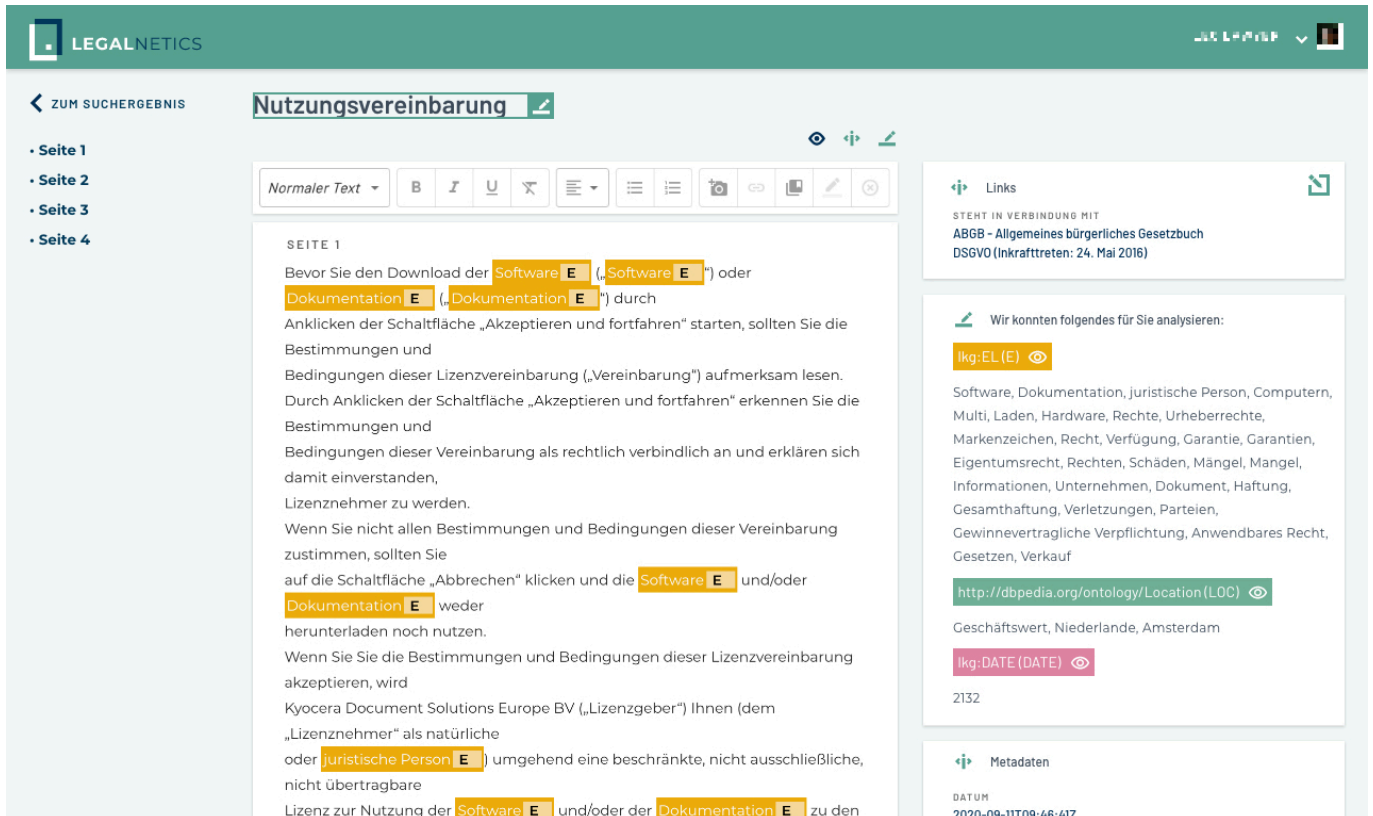


Figure 16 Adding documents to a folder

3.7 UPDATE

Once a document is added to the system it is possible to add / update / delete the annotations of the document. The annotation can be made within the *Editor* which is used to represent the text. The Editor has 3 modes, depending on the document format and user rights:

- Edit: A new document can be created, e.g. a summary to a topic
- Annotate: In this mode the text is protected for modifications, but the annotations can be modified, or new annotation can be added. In this mode it is also possible to add highlights to the text.
- Read: The document and its annotations are protected for modifications



The screenshot displays the LEGALNETICS web application interface. At the top, the title 'Nutzungsvereinbarung' is shown. Below it, a document text is displayed with several terms highlighted in yellow boxes, indicating annotations: 'Software E', 'Dokumentation E', 'juristische Person E', and 'Software E'. The text describes the terms of use for software and documentation. On the right side, there is a sidebar with two sections: 'Links' and 'Metadaten'. The 'Links' section includes a link to 'ABGB - Allgemeines bürgerliches Gesetzbuch' and a link to 'http://dbpedia.org/ontology/Location(L0C)'. The 'Metadaten' section shows the date '2020-09-11T09:46:41Z'.

Figure 17 Manage annotation

In addition, it is possible to provide a short description and a suitable title. Furthermore it is possible to link the document to other documents, legislation and case law.

3.8 EXPORT

For due diligences the system also provides the possibility to export all documents of a folder and its subfolders in the same structure as it was set up in the document store. The title of the document will be the "suitable title" which has been given to the document if exists. If not, the original document name will be used. In the current stage there is no end user front end for this functionality.

3.9 DEMO ACCESS

Access to the pilot can be requested through Cybly homepage: <https://cybly.tech/legalnetics/>

4 ARCHITECTURE

4.1 OVERVIEW

Figure 18 shows new and adapted components developed for this pilot. We will go into detail in the following sections.

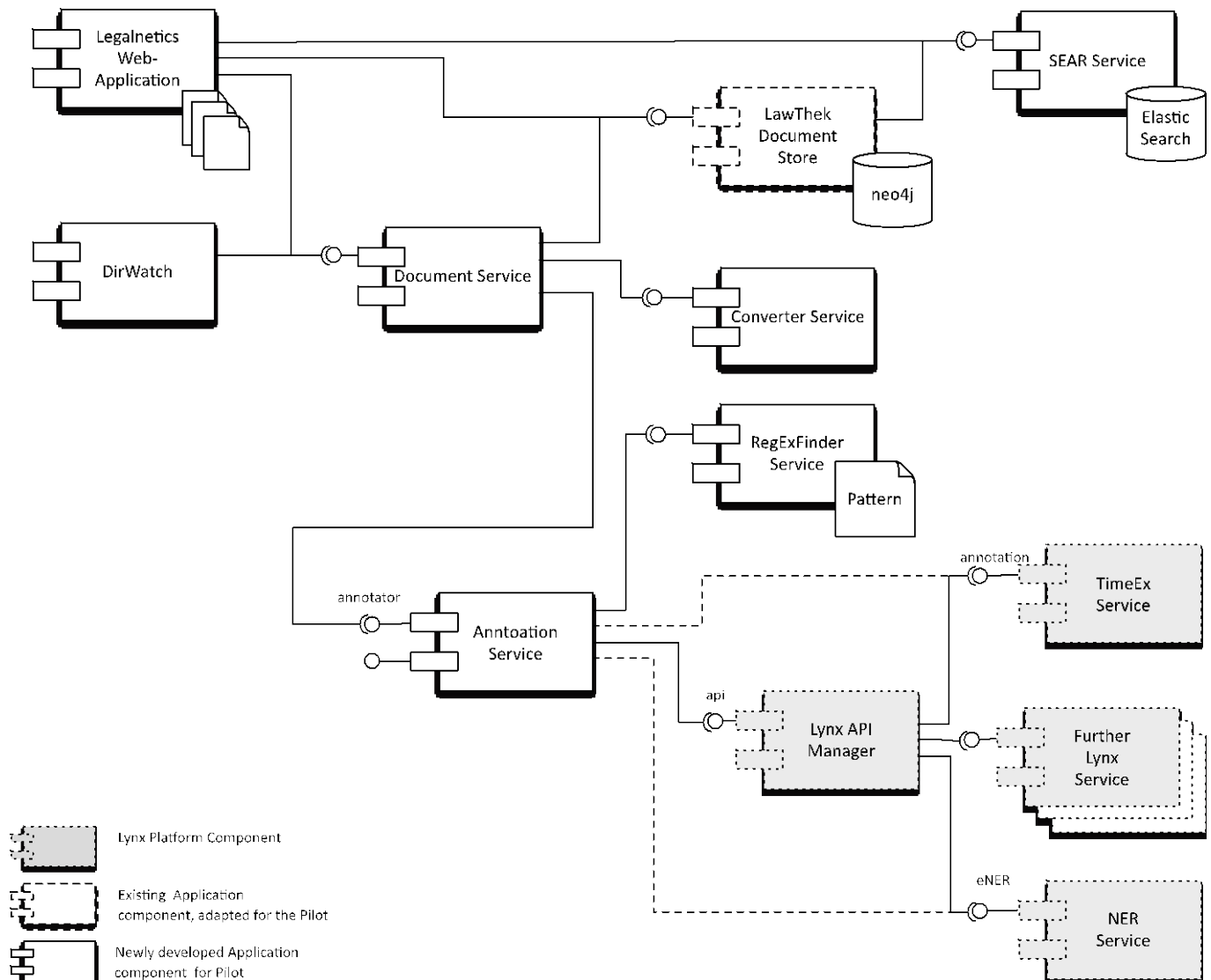


Figure 18. Back-end components

4.2 DIRWATCH

DirWatch is the main service to import contracts into the system. DirWatch is built on top of *fswatch*³. *fswatch* is a command line tool that receives notifications when the contents of the specified files or directories are modified. *fswatch* is available for the 3 major operating systems (Windows, Linux and MacOS) which is important for our use case.

DirWatch can be installed as a service on a dedicated machine to monitor a directory. As there can be a lot of changes and the processing of these events has to be very fast in order to not lose events, every

³ <https://github.com/emcrisostomo/fswatch>

event is imported into a message queue, in our case RabbitMQ⁴. This message queue is then processed by a parallel process to pass the information to the LawThek Document Store.

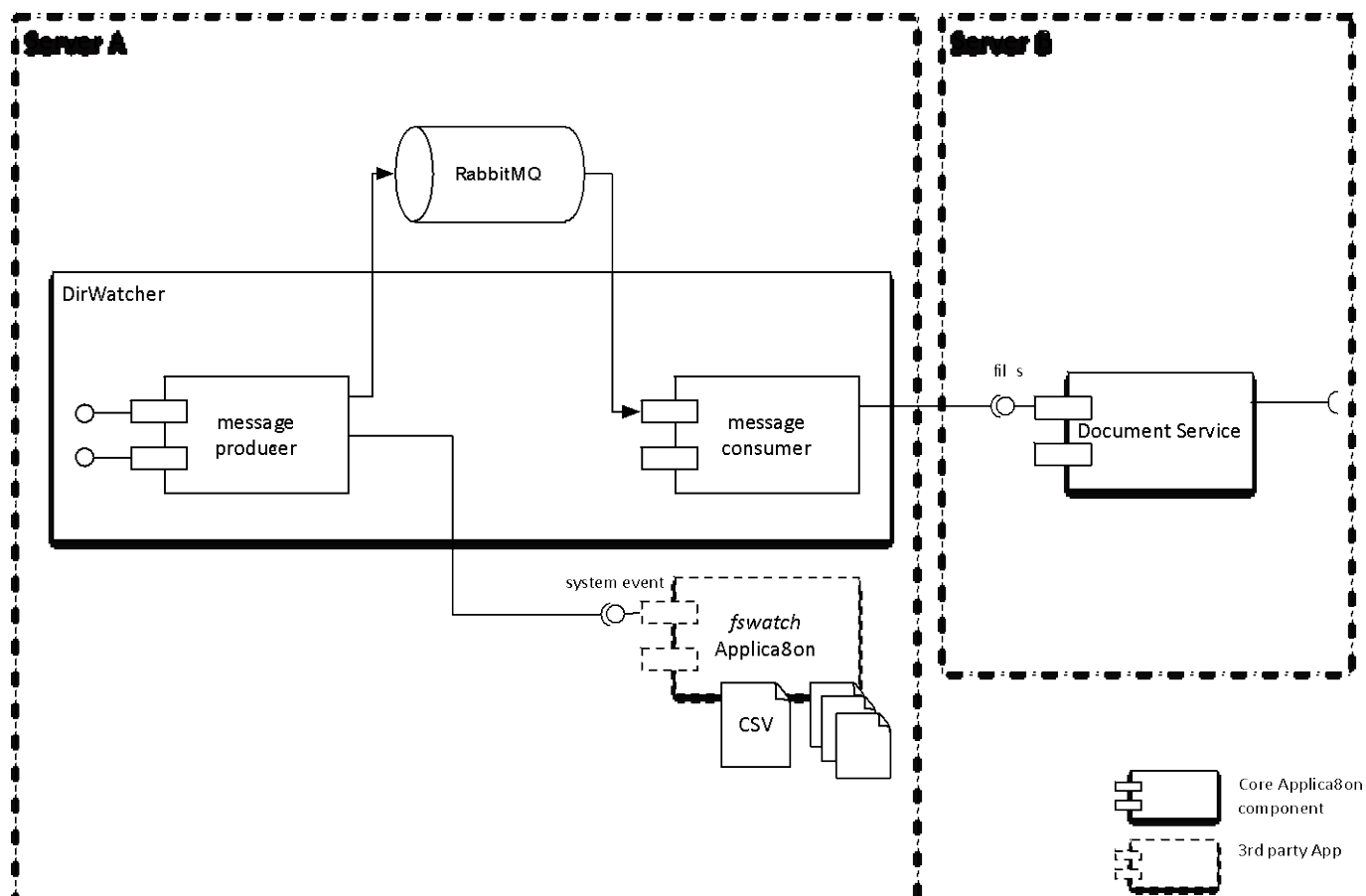


Figure 19. DirWatcher

4.3 DOCUMENT SERVICE

The Document Service is a lightweight workflow application implemented as a Java Spring Boot application. It provides a REST API which performs the following tasks :

- Convert files to a LynxDocument with the Converter Service
- Store, update, delete the file in Customers local LawThek Document Store which is a neo4j database (see Figure 20) for meta data and relationships in combination with file storage to persist the original file
- Store, update, delete the document in the search index of the LawThek Document Store

The Lynx Workflow Manager has not been used for the following reason: The design principals and requirements are different to the workflow here. The workflow within the Document service is static – it is defined as code and is mainly a linear sequence. No additional external services, e.g. database should be used. Resilience within the workflow is not required, as the used services and application already take care of it.

⁴ <https://www.rabbitmq.com>

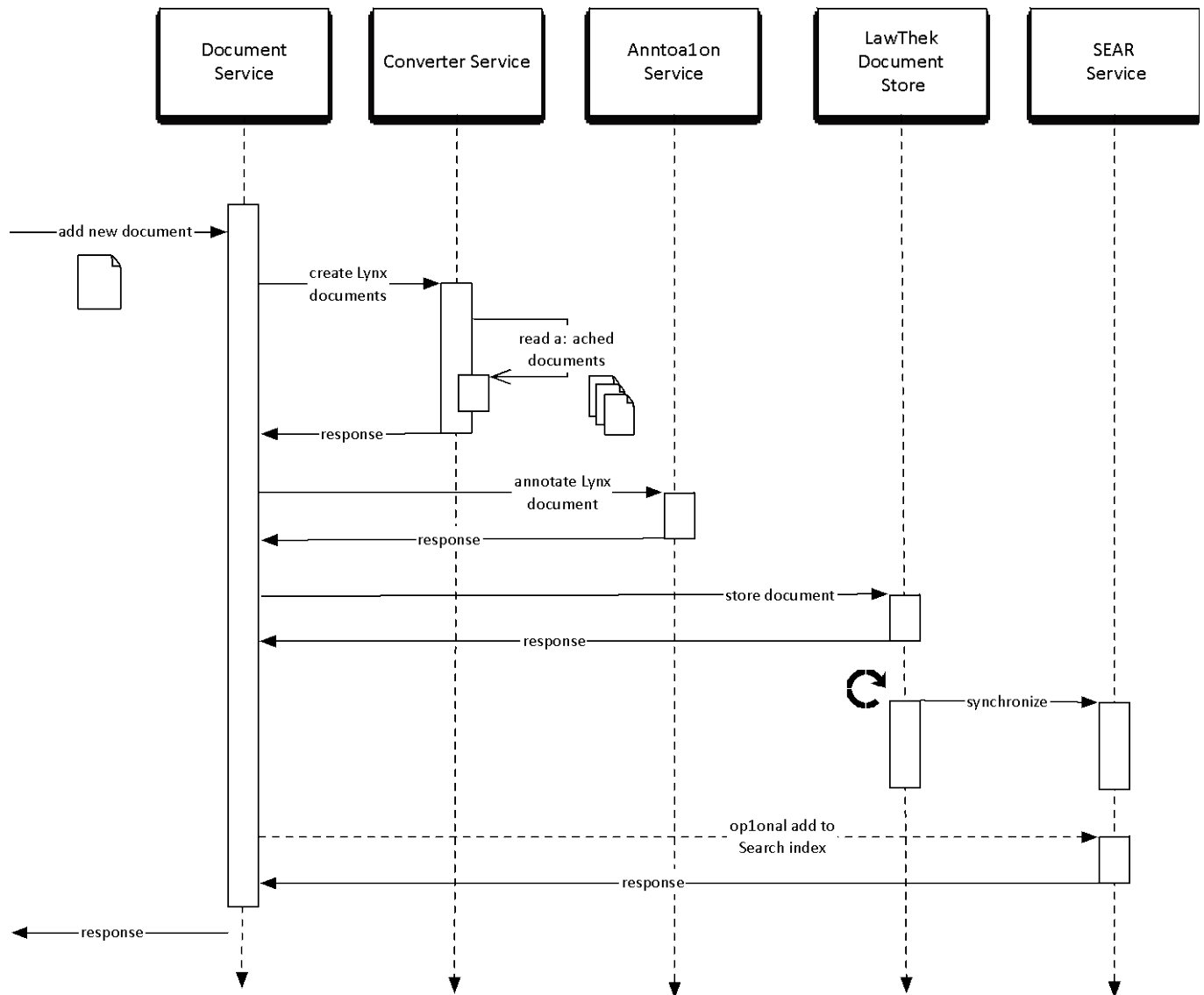


Figure 20. Sequence when a new document is added

4.4 DOCUMENT CONVERTER SERVICE

The Document Converter Service is a Java Spring Boot application. It provides a REST API to extract text content and meta information of different document formats, e.g. e-mail, office documents and pdf files. Afterwards, this information is returned in the form of a Lynx Document enriched with metadata⁵ (in Json-LD format).

In case a document is an e-mail, the attachments are also extracted in a recursion. Finally, a list of Lynx Documents including all e-mails and related attachments are returned. For pdf files an OCR process is performed.

⁵ <http://lynx-project.eu/data2/data-models>

4.5 ANNOTATION SERVICE

The Annotation Service is a Java Spring Boot application which orchestrates the calls to different Lynx Services (Figure 21 Annotation Service) (See also Section 4.7 Lynx Services), and Legalnetics (UseCase) services, and store the result within the LawThek Document Store.

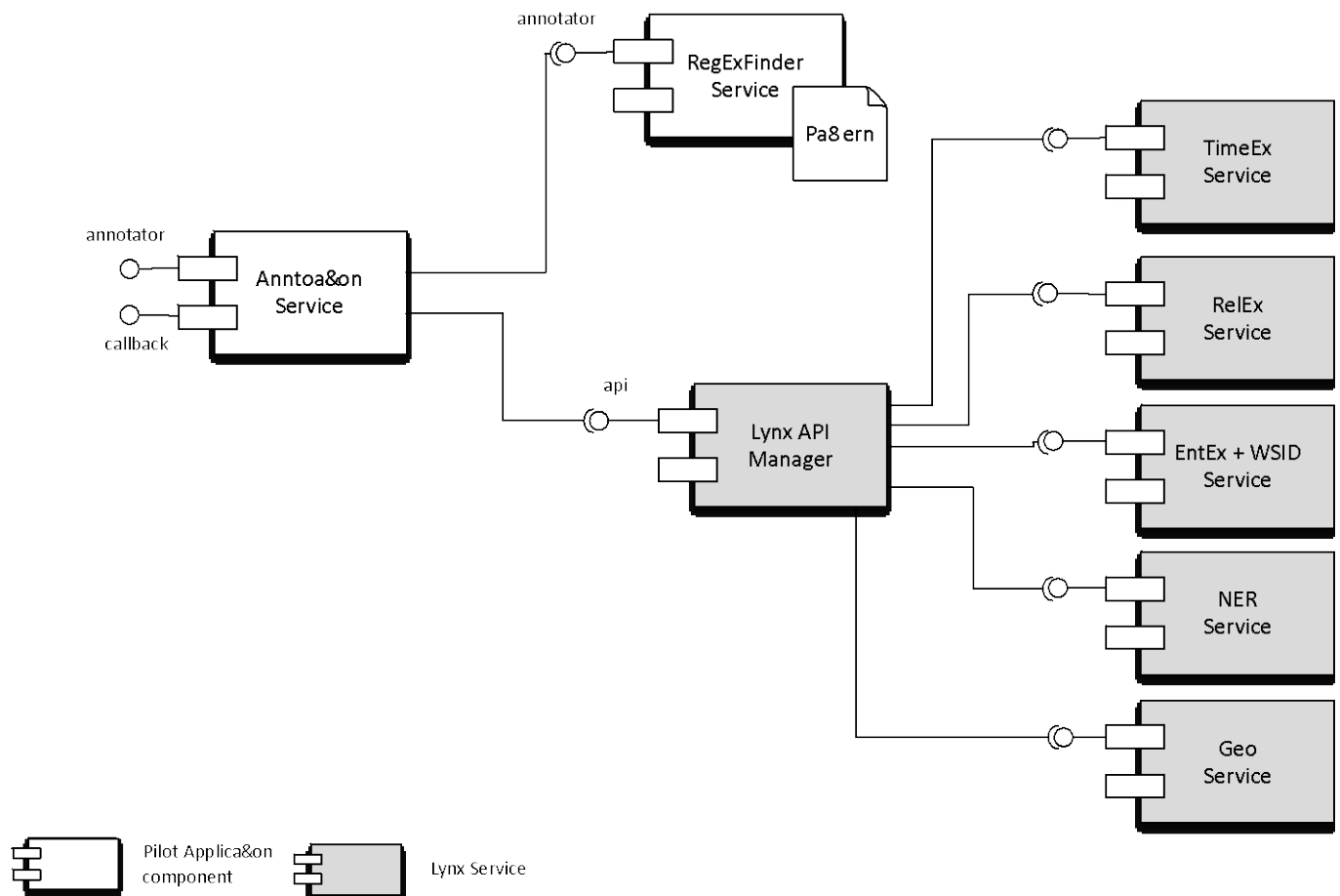


Figure 21 Annotation Service

The RegExFinder Service searches for entities based on Regular Expressions⁶ (Patterns) and passes them back as Lynx annotations. This service is used to extract entities like offer number, customer number, quote number, and consumer price index number of the text corpus.

The regular expression can be defined in a file which is loaded at service start. Figure 22 shows a sample for such a file.

```
{
  "rules":[
    {
      "taClassRef":"lnx:cpi",
      "pattern":[
        "\\(?VPI\\)?\\s?[1|2]\\d\\d\\d\\d\\?"
      ]
    },
    {
      "taClassRef":"lnx:offer-number",
      "pattern":[
        "\\b20[1|0]\\dA\\d\\d\\d\\d\\d\\dB",
        "\\bAN\\d{9}\\dB"
      ]
    }
  ]
}
```

Figure 22 Sample definition for regex pattern

The Lynx Services⁷ used in this pilot (as of the time of editing this document or in a near future) are the following:

- **TimEx** - Temporal extraction. Temporal expressions in documents are detected, e.g., the date when the offer has been made.
- **NER** - Named Entity Recognition – Is used to identify named entities such as persons and organisations (companies).
- **APIM** - Api Manager – It exposes RESTful API to first party clients, end-users and administrators. It will represent the main entry point to the Lynx Services in the future.
- **WM** - Workflow Manager. This service is used to implement the necessary annotation workflows, within Lynx. For the pilot a workflow is defined to perform all annotations, translations and summarizations with a single call to the Lynx Platform and return the information once the result is available.
- **RelEx** - Relation Extraction between entities within a single document will be used to find e.g. cause-effect relation: Example: The agreement ends by 20.1.2020
- **EntEx+WSID** - Entity Extraction and Word Sens Disambiguation Service. The service produces the necessary annotations of the documents enriching the documents with entities from the defined vocabulary.
- **Geo** – to find geographical expressions in documents, mainly addresses.

⁶ https://en.wikipedia.org/wiki/Regular_expression

⁷ <http://lynx-project.eu/doc/api/>

4.8 TECHNOLOGY

The front is developed with Next.js⁸, a JavaScript⁹ library, and uses Material-UI¹⁰ as a component library. It communicates through REST¹¹ calls with the Pilot Back-End System (LegalNetics) and the LawThek¹² system. The LawThek system provides all needed functionalities and includes user management, search, querying individual documents (legislation, case law, summaries and contracts) and the attached meta information including references to other documents.

The back-end system of the pilot is implemented in Java Spring¹³ and follows a microservice architecture like the rest of the LawThek and Lynx Platform. All services are individual applications providing REST interfaces.

⁸ <https://nextjs.org>

⁹ <https://en.wikipedia.org/wiki/JavaScript>

¹⁰ <https://material-ui.com>

¹¹ https://en.wikipedia.org/wiki/Representational_state_transfer

¹² <https://lawthek.eu>

¹³ https://en.wikipedia.org/wiki/Spring_Framework

5 REFLECTIONS AND RECOMMENDATIONS FOR FURTHER DEVELOPMENT

5.1 TEST DATA

A main issue was and is to have test data which can be shared within the consortium. Contracts and contract-related data often include confidential information and also personal data. For personal data we would have to follow the GDPR regulations with the consortium which is not reasonable. For confidentiality we would have needed n-way confidentiality agreements which is also not the case.

Even internally, within Cybly, the document couldn't be shared with the different developers, only selected people are allowed to access the real data. For this reason, often single documents have been created to evaluate the services and to support the development, but when it came to the training of "NLP" services this rapidly leads to problems.

To overcome this, we have investigated with two external parties for anonymization services to anonymize the documents, but this itself would have been a project on its own. It would not have been enough to anonymize the document; we would have to pseudonymize the document again with meaningful data. We are still in contact with these parties, but currently the investigations in this direction are on hold.

5.2 ANNOTATIONS

For training different services annotated data are also needed. Even if there are documents available for testing and training, they are not annotated, to do so we have tools which could be used. The following tools were evaluated / tested:

- Brat¹⁴: Very well-known when showing NLP annotations, but the latest version is from 2012-11-08. To install and configure it is not easy as system requirements are quite old. For this reason, it has been decided not to use this tool.
- doccano¹⁵: A beginner friendly tool which builds on modern technologies, but it is sentence-based. For this reason, this tool couldn't be used.
- INCEpTION¹⁶: The successor of WebAnno. It is a very powerful toolset for annotating text. It offers a lot of functionality and also supports NIF¹⁷ as import format, which is the basis for a Lynx document. Unfortunately, we found this tool at a very late stage, so we have not really used it for a larger document corpus yet. However, we would recommend this tool for the future.

Other small web tools were also evaluated, such as xBatx¹⁸ which has been adapted for a small manual annotation run for the relations extractions (RelEx), but we will no longer use this tool as it has been replaced by the newly developed editor.

Since annotating and correcting existing annotations is an important function within the Use Case, we have now integrated this functionality into the new editor. It allows to create new documents and annotate existing documents (see 3.7 Update).

¹⁴ <https://brat.nlplab.org/>

¹⁵ <https://doccano.herokuapp.com/>

¹⁶ <https://inception-project.github.io>

¹⁷ https://inception-project.github.io/releases/0.17.1/docs/user-guide.html#sect_formats_nif

¹⁸ <https://github.com/xBATx/text-annotation-tool>

5.3 NLP MODELS

As there are too few annotated documents, at least in German, for the training, we had to fall back on standard language models. However, it turned out that the sentence structure in German contracts and the use of abbreviations with a dot "." at the end often leads to the issue that important annotations are not being found. For this reason, we will try to adapt existing language models until the end of the project with the Lynx Consortium partners in such a way that they can recognize the German legal syntax and the hit rate is increased.

6 REFERENCES

- Flitsch, Martina (2010) Verträge und Vertragsmanagement in Unternehmen.
- Hamming, Richard W. (1962). Numerical Methods for Scientists and Engineers. New York: McGraw-Hill.; second edition 1973